ED 308 204 TM 013 414

AUTHOR Livingston, Samuel A.; And Others
TITLE What Combinatior. of Sampling and Equating Methods Works Best? Revised.
SPONS AGENCY Educational Testing Service, Princeton, N.J.
PUB DATE 26 Apr 89
NOTE 27p.; Revised version of a paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, March 27-31, 1989).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Comparative Analysis; Data Collection; *Equated Scores; High Schools; *High School Students; Latent Trait Theory; *Research Methodology; *Sampling; Statistical Analysis; Test Bias; Test Format
IDENTIFIERS Equipercentile Equating; Frequency Estimation Equipercentile Equating; Large Scale Programs; Levine Equating Method; Scholastic Aptitude Test; Three Parameter Model; Tucker Common Item Equating Method

ABSTRACT

Combinations of five methods of equating test forms and two methods of selecting samples of students for equating were compared for accuracy. The two sampling methods were representative sampling from the population and matching samples on the anchor test score. The equating methods were: (1) the Tucker method; (2) the Levine method; (3) the chained equipercentile method; (4) the frequency estimation; and (5) an item response theory (IRT) method; specifically, the three-parameter logistic model. The tests were the verbal and mathematics sections of the Scholastic Aptitude Test. The criteria for accuracy were measures of agreement with an equivalent-groups equating based on more than 115,000 students taking each form. Much of the inaccuracy in the equatings could be attributed to overall bias. The results for all equating methods in the matched samples were similar to those of the Tucker and frequency estimation methods in the representative samples; these equatings made too small an adjustment for the difference in the difficulty of the test forms. In the representative samples, the chained equipercentile method showed a much smaller bias. The IRT and Levine methods tended to agree with each other and were inconsistent in the direction of their bias. Five tables and four figures present study data. (Author/SLD)

# WHAT COMBINATION OF SAMPLING AND EQUATING METHODS WORKS BEST?[1,2]

Samuel A. Livingston
Neil J. Dorans
Nancy K. Wright

Educational Testing Service

Revised April 26, 1989

[1] Paper presented in a Symposium entitled Selecting Samples for Equating: To Match or Not to Match, at the annual meeting of the American Educational Research Association, San Francisco, March, 1989.

[2] The authors thank Karen Carroll, Miriam Feigenbaum, Nancy Feryok, and Judy Fisher for performing the data analyses reported in this paper. The opinions expressed in this paper are those of the authors and do not represent official positions of Educational Testing Service. Funding for this and other papers in this symposium was provided by the Program Research Planning Council of Educational Testing Service.

## Abstract

Combinations of five methods of equating test forms and two methods of selecting samples of students for equating were compared for accuracy. The two sampling methods were representative sampling from the population and matching samples on the anchor test score. The equating methods were the "Tucker", "Levine", "chained equipercentile", "frequency estimation", and IRT (3PL) methods. The tests were the verbal and mathematical sections of the Scholastic Aptitude Test. The criteria for accurac' were measures of agreement with an equivalent-groups equating based on more than 115,000 students taking each form.

Much of the inaccuracy in the equatings could be attributed to overall bias. The results for all equating methods in the matched samples were similar to those for the Tucker and frequency estimation methods in the representative samples; these equatings made too small an adjustment for the difference in the difficulty of the test forms. In the representative samples, the chained equipercentile method showed a much smaller bias. The IRT and Levine methods tended to agree with each other and were inconsistent in the direction of their bias.

# WHAT COMBINATION OF SAMPLING AND EQUATING METHODS WORKS BEST?

Samuel A. Livingston, Neil J. Dorans, and Nancy K. Wright
Educational Testing Service

When a testing organization equates two forms of a test, the statisticians often have a choice of ways to select samples of student test papers to use in the equating. One possibility is simply to use all available test papers, but this choice may not always be the best choice. The statisticians also have a choice of methods to use in estimating the equating relationship between the two forms of the test. What combination of sampling and equating methods works best?

The present study was an attempt to answer this question for a particular type of equating situation that is common in large-scale testing programs. In this situation a form of a test being given for the first time -- the "new form" -- is equated to a form of the test that was given previously but is no longer being given -- the "old form". The two forms are linked through an "anchor test" that was administered to students taking the new form and to students taking the old form. The anchor test may consist of a set of items contained in both forms, or it may be a separate test of similar abilities. The group of students taking the new form may or may not be similar in ability to the group that took the old form.

This study is based on the assumption that there is a population of students for whom the equating is intended to be correct -- the "target population" -- and that it is possible to draw a sample of the students taking the new form in such a way that this "new form sample" will be representative of the target population. A second important assumption is that the "true" equating relationship -- the one to be estimated, as closely as possible, from the available data -- is the equipercentile relationship in the target population. This equating will be referred to as the "target equating". It is the equipercentile equating that would result if the entire target population took both forms of the test, with no practice or sequence effects.

## The Sampling and Equating Methods

The present study was a comparison of several combinations of two sampling methods and five equating methods. Both sampling methods assume that the new form sample -- the sample of students taking the new form whose test papers are used in the equating process -- is representative of the target population. The two sampling methods differ in the way the "old form sample" is selected. The first sampling method simply chooses the old form sample randomly (or by a quasi-random procedure such as "spaced sampling") from the population of students who took the old form. This method will be referred to as "representative" sampling, although the resulting samples are only approximately representative of their parent populations. The second sampling method uses the anchor test score as a stratifying variable to match the old form sample to the new form sample. It guarantees that the old form sample and the new form sample will have the same distribution of scores on the anchor test. This method will be referred to as "matched" sampling.

The five different equating methods compared in the present study (see Dorans, 1989, for further details) include two linear methods, i.e., methods that assume that the equating relationship can be represented on a graph by a straight line. The other three methods are curvilinear methods; they make no such assumption. The first of the two linear methods will be referred to as the Tucker method (Angoff, 1984, pp. 109-112). It equates the estimated means and standard deviations of the scores that would have been observed if the students in the old form sample and those in the new form sample had taken both forms of the test. It is based on the assumption that the linear regression of new-form score on anchor-test score -- slope, intercept, and residual variance -- is the same in the old form sample (where it is unobserved) as in the new form sample (where it is observed). It makes a similar assumption for the regression of new-form score on anchor-test score.

The second linear equating method will be referred to as the Levine method (Angoff, 1984, pp. 113-115). This method is similar to the Tucker method, except that the assumptions used to estimate the means and standard deviations in the combined sample are not the same as those for the Tucker method. The assumptions for the Levine method involve the regressions of new-form and old-form true scores on anchor-test true scores. Also, the variance of errors of measurement on each form (rather than the residual variance in the regressions) is assumed to be the same in both samples.

The third equating method is based on a procedure called "frequency estimation" (described in Angoff, 1984, p. 113). This procedure estimates, for each form, the joint distribution of scores on that form and the anchor test. This joint distribution is estimated for a group of students with a specified distribution of scores on the anchor test; we typically use the distribution in the combined (old form and new form) sample.[1] The key assumption is that the conditional distribution of scores on the new form, given the score on the anchor test, is the same in the old form sample (where it is unobserved) as in the new form sample (where it is observed). The method makes a similar assumption for the old form. Summing over scores on the anchor test yields estimated distributions of scores of the combined sample on the new form and on the old form. The third method included in this study was an equipercentile equating of these estimated distributions.

The fourth equating method (Angoff, 1984, p. 116) is the composite of two equipercentile equatings: an equating of the new form to the anchor test in the new form sample and an equating of the anchor test to the old form in the old form sample. Marco, et al. (1983) referred to this method as the "direct equipercentile" method. We prefer the term "chained equipercentile" method, because it consists of two separate equipercentile equatings, linked by the anchor test.

[1] This version of the method is consistent with Angoff's description. When only one of the samples is representative of the target population, it makes sense to use the anchor score distribution of that sample, rather than the combined sample. Nevertheless, in this study we have used the method as described by Angoff.

The fifth method is based on item-response-theory (IRT), specifically, the three-parameter logistic model (Lord, 1980). This method began with a simultaneous ("concurrent") calibration of all the items in both forms, using the computer program LOGIST (Wingersky, et al, 1982). The estimated item parameters resulting from this calibration were used to estimate the expected scores on the two forms at several closely spaced levels of ability. The resulting (x,y) pairs define the equating transformation. For conciseness, this method will be referred to simply as "IRT".

When the new form and old form samples of students have identical score distributions on the anchor test, both linear methods described above become identical to a simple linear equating of the observed sample means and standard deviations. These methods use the anchor test scores to adjust for ability differences between the new form and old form samples. When the two samples have identical score distributions on the anchor test, there are no adjustments to be made. Similarly, in perfectly matched samples, the chained equipercentile method and the frequency estimation equipercentile method both become equivalent to a simple equipercentile equating of the observed distributions (except for small differences introduced in the interpolation procedure). Therefore, to the extent that perfect matching is possible, the six methods described above are reduced to three methods in the matched samples: linear equating, equipercentile equating, and IRT equating.

## The Data

The tests equated in this study were the verbal and mathematical portions of two forms of the Scholastic Aptitude Test (SAT) that had been "spiraled" (i.e., administered in alternating sequence) in a regular SAT administration involving approximately a quarter of a million students. The equating of the SAT excludes students not in their junior or senior year of high school, and these students' papers were excluded from the equatings in this study. The remaining 236,000 students are the target population. The spiraling of test forms divided this population into groups of 119,000 and 117,000 students. (One group is slightly larger because of the way the test booklets are packaged.) A sample of 117,000 or more students, sampled by spiraling test forms, can be assumed to represent very closely the ability distribution of the full population. Therefore, the equipercentile equating of the score distributions of these two groups of students can be assumed to be, for all practical purposes, the equating relationship in the target population.

One of the two forms had been designated as the "new form" and the other as the "old form" in the equating that had been used to report scores on these two forms, and these designations were kept in conducting the study. No anchor test was necessary for equating these two forms of the SAT to each other. However, several anchor tests were administered with these two forms for the purpose of equating them to other (past and future) forms of the SAT. These anchor tests were "spiraled" in the population of test-takers, so that each combination of test form and anchor test was taken by a stratified random sample of approximately 8,000 students. Four of these anchor tests -- two verbal and two math -- were administered with both of the forms in this study. The correlation between the anchor score and the score to be equated varied from .86 to .88.

The four anchor tests made it possible to create, artificially, several anchor equating situations in which the populations of students taking the old form differed systematically in ability from the populations taking the new form -- situations in which the true equating relationship in the target population was known. Each equating situation consisted of a pair of artificial pseudo-populations linked by an anchor test. The new form pseudo-population in each pair was simply the entire group of students taking the new form and the anchor test. Each old form pseudo-population was selected to be of systematically lower ability than the corresponding new form pseudo-population. The old form pseudo-population in each pair was selected from the students taking the old form and the anchor test, by removing a portion of the higher-ability students. The old form pseudo-populations for equating the verbal test were selected on the basis of their math scores, and vice versa, to avoid selecting on either the anchor score or the score to be equated.

Each new form pseudo-population, was paired with two different old form pseudo-populations of different ability levels. One of the old form pseudo-populations was selected to have a mean ability level approximately 0.2 standard deviations lower than the new form pseudo-population. This pseudo-population will be referred to as the "0.2 population". The other pseudo population was selected to have a mean ability level approximately 0.4 standard deviations lower than the new form pseudo-population. This pseudo-population will be referred to as the "0.4 population".[2] The "0.2 populations" varied in size from 6148 to 6658 students; the "0.4 populations" varied in size from 4367 to 4887 students.

Although the new form and (particularly) the old form pseudo-populations in this study are artificial, they are made up of real students. The data are not simulated data. They are the actual test responses of real students sitting in testing rooms with their Number 2 pencils, trying to get into the colleges of their choice.

## Samples for Equating

In every equating in this study, the new form sample was a representative sample of the new form pseudo-population. These samples were drawn by a spaced random sampling procedure: dividing the data file into equal-sized blocks of a specified size and selecting a specified number of students randomly from each block. The "representative" old form samples

[2] The correlation between verbal and math scores is approximately .70. A "0.2 population" for equating verbal scores was selected by specifying a distribution of math scores that had a mean (0.2 /.70) standard deviations below that of the full old form population. The resulting "population" had a mean verbal score approximately 0.2 standard deviations below that of the full population. A similar procedure was used for selecting the other old form "populations".

were drawn in the same way from their respective old form pseudo-populations. All the representative samples were drawn to include approximately 3000 students.

The "matched" old form samples were to be drawn from their respective populations by using the anchor test as a stratifying variable and randomly selecting the same number of students at each score level as were in the new form sample. However, it was necessary to modify this plan, because there were not enough high-ability students left in the old form pseudo-populations, particularly the "0.4 populations". In selecting matched samples from the "0.2 populations" it was possible to select samples with anchor score distributions very similar to (though not exactly the same as) those of the new form samples. In selecting matched samples from the "0.4 populations" it was necessary to reduce the sample size proportionally, from 3000 to approximately 1500, and even then the samples were not perfectly matched on the anchor test score.

The design of the study therefore involved a total of 16 pairs of samples for equating. The design is illustrated in Figure 1. Table A1, in the appendix, summarizes the characteristics of the equating samples: the type of sample (representative or matched and, for the old form samples, which population the sample was selected from), the number of students in the sample, and their mean score on the anchor test. If the "matched" old form samples were perfectly matched to the new form samples, their mean anchor scores would be exactly equal to those of the corresponding new form samples. As Table A1 shows, they were not.

## Criteria for Accuracy

The main criterion for judging the overall accuracy of each equating was the root-mean-squared deviation (RMSD) of the equated scores of the full new form population from their equated scores determined by the target equating. The RMSD is computed by the formula

$$\text{RMSD} = \surd\ [\ (\ \Sigma\ n(x)\ [\hat{y}(x) - y(x)]^2\ )\ /\ \Sigma\ n(x)\ ]\ ,$$

where n(x) is the number of examinees (in this case, the number of juniors and seniors in the full population) with raw score x on the new form, y(x) is the corresponding exact scaled score on the old form as determined by the criterion equating, and $\hat{y}(x)$ is the corresponding exact scaled score on the old form as determined by the other equating to be compared with the criterion equating. The summation is over the raw score levels on the new form. The equated scores are expressed on the College Board 200-to-800 scale, and the RMSD statistics are in terms of this scale. Since the scores are reported in ten-point intervals, an RMSD statistic of 3.3 for an equating can be interpreted to mean that the equated scores of the new form population are, on the average, about one-third of a score level away from what they shou_d have been. The standard deviation of scaled scores in the f 'l test-taker population, for both the verbal and math scores, is about 100 points; adjacent score levels (e.g., 450 and 460) differ from each other by about one-tenth of a standard deviation. We consider an RMSD statistic of 5 or more as an indication of a problem equating.

A secondary criterion for judging the accuracy of each equating was its bias -- its tendency to produce equated scores that were systematically too high or too low. The overall bias statistic is an average value for the new form population. The bias of the equating is computed by the formula

$$\text{Bias} = ( \Sigma\, n(x)\ [\hat{y}(x) - y(x)]\ ) \;/\; \Sigma\, n(x)\ ,$$

where the symbols have the same meaning as in the formula for the RMSD. Note that negative bias in one portion of the score range will cancel out positive bias in another portion of the score range. Therefore, the bias statistic is not a good basis for evaluating an equating unless all the equated scores are too high or all are too low. However, the bias statistic is always valuable as a diagnostic tool for investigating the reason for a large RMSD statistic.

## Results

Table 1 shows the mean and standard deviation, in the new form population, of the scaled scores that would result from each of the equatings. Table 2 shows the bias and RMSD of the scaled scores resulting from each of the equatings. These statistics are expressed in the units of the SAT 200-to-800 scale. The target equating is the equipercentile equating in the full population. The information in Table 2 is presented graphically in Figures 2a to 4d.

Figures 2a to 2d compare the accuracy of eight combinations of sampling and equating methods in the four equating situations using the "0.4 population" as the old form population. Each of these four plots contains eight data points; each data point represents a different combination of sampling and equating methods. The accuracy of the equating, as indicated by the RMSD statistic, is represented by the diagonal distance from the data point to the origin. The horizontal component of this distance represents the bias in the equating. The vertical component has no simple interpretation; it represents all the other factors (i.e., other than a constant bias) that contribute to the RMSD.

All four plots in Figures 2a to 2d show a similar clustering of the data points. At the left of the graph, indicating a large negative bias, are the data points for the Tucker method and the frequency estimation method in the representative samples and for all three methods (linear, equipercentile, IRT) in the matched samples. Toward the center of each figure, indicating less negative bias, is the data point for the chained equipercentile method in the representative samples. At the right of each figure, indicating the least negative bias or, in two cases, a positive bias, are the data points for the Levine method and the IRT method in the representative samples.

Figures 3a to 3d show the same information as Figures 2a to 2d, but for the equatings in which the old form samples were drawn from the "0.2 populations". The data points in these four plots are all closer to the origin, indicating more accurate equatings. They tend to show the same general pattern as those in Figures 2a to 2d, but there are some differences. Figure 3b shows the matched-sample methods clustering separately from the Tucker and frequency estimation methods in the representative samples. The reason for this separation appears to be that, because of sampling variability, the relationship between the anchor test scores and the full verbal scores was not the same in the representative sample as in the matched sample.[3]

There is one result that appears consistently in the equatings of the math scores but not in the equatings of the verbal scores: although all methods in the matched samples show a consistent negative bias, the IRT method shows somewhat less bias than the other methods. In the equatings of the verbal scores, the IRT equating in the matched samples generally shows a smaller RMSD than the other matched-sample methods, but about the same degree of bias.

Figures 4a to 4d show the bias and RMSD statistics for equatings done by each method in the full subpopulations, i.e., all students taking each combination of test form and anchor test. These statistics represent the best results that can reasonably be expected from each method. Any bias in these equatings is the result of sampling variability in the full-test and anchor-test scores of the subpopulations. These methods show the two linear methods clustering closely together with the two equipercentile methods, while the IRT method tends to produce somewhat different results, especially for the equatings of the math scores.

Note that none of the anchor equatings in Figures 4a to 4d exactly reproduces the target equating. Also note that all four sets of equatings show some degree of bias, especially the equatings of the verbal scores through anchor "vb". This bias is a result of sampling variability in the full test and anchor scores. For example, the subpopulation taking the old form and anchor test "vb" did particularly well on the full test, averaging about 0.37 raw-score points (.024 SD) better than the group of all students taking the old form. The subpopulation taking the new form and anchor "vb" averaged only 0.06 raw-score points (.004 SD) better than the group of all students taking the new form. Yet, this difference was not reflected in the anchor scores; the anchor score means of these two groups differed by only .002 SD. As a result, all the equatings in these two subpopulations were

[3] For the equating results shown in Figure 3b, the mean difference between the anchor scores of the matched old-form sample and the representative old-form sample was .18 SD, while the mean difference in their full verbal scores was only .12 SD. In Figure 3a, which showed no such separation between the matched-sample results and the Tucker and frequency estimation results, the corresponding mean differences were .19 SD for the anchor test scores and .17 SD for the full verbal scores.

biased in a positive direction. For this reason it may be useful to recompute the bias and RMSD statistics for each equating, with the corresponding subpopulation equating as the target equating.

Table 3 shows this comparison. Each of the bias and RMSD statistics in Table 3 compares the sample equating with its own target equating -- one that uses the same equating method and the same anchor test. Each of these special target equatings was determined in the subpopulation of all students taking the particular anchor test. That is, the subpopulation equatings of Figures 4a to 4d become the target equatings in Table 3. However, the bias and RMSD statistics for each of these comparisons are computed for the full population, not the subpopulation. These results show the same consistent bias for the Tucker and frequency estimation methods and for all methods applied in the samples matched on the anchor test. The results for the other methods are less clear. In the representative samples, the chained equipercentile method consistently produced equated scores that were lower than those produced by the Levine and IRT methods, even when each method is compared against a target equating by the same method. It is difficult to say whether this difference reflects a negative bias in the chained equipercentile method, a positive bias in the Levine and IRT methods, or both.

## A Partial Explanation

The clustering of the methods in the results of this study is not a coincidence. The matched-sample methods tended to cluster together because they all use the information contained in the anchor scores in the same way: to create matched samples of test-takers. Once the samples have been selected to have identical anchor score distributions, there is no relevant information remaining to be extracted from the anchor scores. Therefore, all methods tend to produce similar results in the matched samples.

The Tucker and frequency estimation methods tend to cluster with the matched-sample methods because they use the anchor score in a similar way: as a conditioning variable. Both these methods assume, for equating purposes, that if two groups of test-takers have identical score distributions on the anchor test, they should have identical distributions of equated scores. Putting it another way, students with the same anchor score are assumed to be exchangeable between the old form and new form populations. If either of these methods is applied to samples with identical anchor score distributions, the method assumes, in effect, that the samples are of equal ability and simply equates the observed percentiles or the observed means and standard deviations. These methods can be understood as attempts to estimate the equating relationship in samples matched on the anchor test. Consequently, their results tend to agree with the results of equatings actually performed in samples matched on the anchor test.

This explanation accounts for the tendency of these methods to show a negative bias in this study. The old form pseudo-populations in this study were systematically less able than the new form population. The anchor score is an imperfect measure of the abilities measured by the old form and new form scores. When samples from different populations are matched on the anchor test, there is a "regression effect", so that matching on the anchor score does not completely remove the ability difference between the samples. If the old form population was less able than the new form population, the old form sample will still be somewhat less able than the new form sample, even though their anchor scores do not show any difference. Since the students in the old form sample are less able than they would be if they were truly matched on ability, they tend to earn lower scores on the old form (i.e., lower than they would if they were truly as able as the new form sample.) Therefore, the old form appears relatively harder, the new form appears relatively easier, and the equating does not award a high enough equated score for a given raw score on the new form. As a result, the equating is biased in a downward direction. (Ironically, in the equatings shown in Figure 3b, the inconsistency between the full-test and anchor-test scores of the "representative" old-form sample tended to counteract this general tendency. Consequently, the results of the Tucker and frequency estimation methods in the representative samples were free of bias in this one case.)

It is difficult to explain fully why the Levine method and the IRT method tended to agree so consistently in the matched samples. A comparison of the actual conversion lines also showed these two methods tending to deviate from the target equating in the same direction in the same parts of the score scale. The two methods are similar in that they both assume exchangeability (across old-form and new-form samples) for students with the same true score, rather than for students with the same observed anchor score. This assumption seems to produce a general agreement in the results of these two methods, even though they differ in their other assumptions, in the type of data they use (item-level, or score-level), and even in their definition of equating!

It is not surprising that the chained equipercentile method did not cluster with any of the other methods. This method is based on an entirely different logic from the other methods; it uses the information in the anchor scores in a different way. It does not estimate old-form and new-form distributions on the basis of assumptions about the exchangeability of students between old and new form samples. Instead, it simply equates the new form to the anchor in the new-form sample and the old form to the anchor in the old-form sample. The implicit assumption is that the equating relationship between each form and the anchor test is the same in the group where it is unobserved as it is in the group where it is observed. That is, the chained equipercentile method assumes that equating relationships are stable across populations of test-takers. This method tended to produce surprisingly good results, particularly in its lack of bias. The weakness of this method was that it tended to produce score conversion lines that fluctuated greatly around the target equating. Therefore, it seems reasonable to expect that the addition of a smoothing step would substantially reduce the RMSD for this method, possibly making it the best of the methods tested in this study.

## Limitations of the Study

In any study such as this one, the generality of the findings are subject to question. While our results are similar to those of Marco, et al. (1983), we cannot be sure that they would generalize to all tests equated through an anchor. Although this study represented a considerable amount of effort by several people, its scope is limited. It involved only two tests and only one population of test-takers. Each of the pseudo-populations was only one of many that could have been selected by the specified procedure. Each equating sample was only one of many that could have been drawn from its parent pseudo-population. Would we have found the same results in equating a test through an internal anchor? What if the populations had been selected on a variable that correlated much greater or much less than .70 with the scores to be equated?

The subpopulations of students taking the different combinations of test form and anchor test were not perfectly representative of the full population. Their full-test scores and anchor scores showed differences that would not have occurred if the subpopulations had been perfectly representative of the full population. Differences such as these might be expected because of sampling variability. Nevertheless, they tended to introduce bias into the equatings. As a result of these differences, even the equatings in the subpopulations were not free of bias. The amount and direction of bias varied, as might be expected, from one pair of subpopulations to another.

A different type of limitation of the study is the way in which the pseudo-populations were created. The method we used -- selecting on a variable correlated .70 with the scores to be equated -- may or may not correspond to the way ability differences between old-form and new-form populations arise in the real world of educational testing.

## Implications for Research and Practice

One promising area for further investigation concerns the many equating methods and variations of methods that we did not investigate. How much better would the chained equipercentile method have performed if we had smoothed the distributions before equating? What would we have found if we had used a different IRT parameter estimation method?

Another promising area for investigation concerns other possible sampling methods. The anchor test is not the only possible variable to use for matching equating samples. Would matching samples on some other variable produce better equating results than simply choosing samples that are representative of their respective populations? The Levine and IRT equating methods assume that the best matching variable to use, if it were possible to do so, would be the student's true score. However, it is not possible to select students on the basis of their true scores. Ideally, the right variable to use as a basis for matched sampling would be a measure of whatever is causing the old-form and new-form populations to differ systematically in ability. In practice, we cannot even know what this variable is; we certainly cannot measure it. However, we can do the next best thing. Applying a statistical technique originally developed for medical research, we can select on a "propensity score" (Rosenbaum and Rubin, 1985). A propensity score is similar to a discriminant function; it is the linear combination of all the variables we can measure that best discriminates between the two populations, in this case, the old-form and new-form populations. How accurately could we equate test forms if we selected samples by matching on a propensit· score?

In addition to ...ese practical research questions, there are other types of studies that could improve our understanding of the effects of sampling on equating results. Suppose we repeated this study with samples matched on the variable originally used to construct the pseudo-populations (i.e., math scores for equating the verbal test, and vice versa). Would the equating results be unbiased? If we used this variable as an anchor in the Tucker and frequency estimation equatings, would the results be unbiased? If we constructed new pseudo-populations by selecting on the anchor test itself, would matching on the anchor score produce unbiased equatings?

Although additional research is desirable, those who have the responsibility for equating test scores cannot wait for additional results before deciding which sampling and equating methods to use. What guidance, if any, do the results of this study offer them in making these decisions? When populations differ in ability, matching samples on the anchor test does little to enhance the accuracy of the equating. It makes little difference in the results of the Tucker and frequency estimation methods, and it tends to make the results of any other equating methods very similar to those of the Tucker and frequency estimation methods. A better way to deal with the ability difference would be to select a representative sample from each population and use an equating method that does not assume exchangeability for test-takers classified by their anchor test scores. An alternative would be to select samples by matching on something other than the anchor test, possibly combining several such variables into a single "propensity score".

If matching on a propensity score is the right thing to do, why even bother to study the effects on matching on the anchor test score? The reason is that matching on a propensity score is a much more laborious process than matching on the anchor score. If matching on the anchor test would produce good results, the more involved procedure of matching on a propensity score would be superfluous. In the present study, matching on the anchor test did not produce good results. It appears that selecting equating samples by matching on a propensity score may be a technique worthy of further investigation.

## References

Angoff, W. H. (1984) *Scales, Norms, and Equivalent Scores*. Princeton, NJ: Educational Testing Service.

Dorans, N. J. (1989, March) *The equating methods and sampling designs*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Lord, F. M. (1980) *Applications of Item Response Theory to Practical Testing Problems*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Marco, G. L., Petersen, N. S., and Stewart, E. E. (1982) A test of the adequacy of curvilinear score equating models. *New Horizons in Testing*, 147-177.

Rosenbaum, P. R. and Rubin, D. R. (1985) Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *American Statistician*, *39*, 33-38.

Wingersky, M. S., Barton, M. A., and Lord, F. M. (1982) *LOGIST User's Guide, LOGIST 5, Version 1*. Princeton, NJ: Educational Testing Service.

Table 1. Mean equated scores (on SAT 200-to-800 scale) of full new-form population using raw-to-scale score conversion from each equating.

| | | Sampling method: | Matched on Anchor | | | Representative | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Equating method: | Linear | Equi% | IRT | Tucker | Freqency est. | Chained equi% | Levine | IRT |
| Old form sample | Anchor test | | | | | | | | | |
| 0.4 | va | mean | 416 | 415 | 416 | 415 | 416 | 420 | 423 | 423 |
| | | SD | 103 | 103 | 105 | 101 | 102 | 104 | 106 | 107 |
| 0.2 | va | mean | 421 | 421 | 421 | 421 | 421 | 423 | 425 | 424 |
| | | SD | 104 | 104 | 106 | 103 | 104 | 105 | 106 | 107 |
| 0.0 | va | mean | | | | 424 | 424 | 423 | 424 | 424 |
| | | SD | | | | 105 | 105 | 105 | 105 | 106 |
| 0.4 | vb | mean | 420 | 420 | 420 | 419 | 420 | 425 | 428 | 428 |
| | | SD | 103 | 103 | 105 | 102 | 103 | 106 | 108 | 109 |
| 0.2 | vb | mean | 422 | 422 | 422 | 425 | 425 | 427 | 428 | 429 |
| | | SD | 104 | 104 | 105 | 102 | 102 | 104 | 105 | 106 |
| 0.0 | vb | mean | | | | 427 | 426 | 427 | 427 | 427 |
| | | SD | | | | 105 | 105 | 105 | 105 | 106 |

(Target equating for verbal scores: mean = 425; SD = 105)

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 0.4 | ma | mean | 464 | 464 | 465 | 465 | 466 | 472 | 476 | 477 |
| | | SD | 116 | 117 | 118 | 116 | 117 | 118 | 121 | 121 |
| 0.2 | ma | mean | 471 | 471 | 472 | 470 | 470 | 473 | 474 | 476 |
| | | SD | 116 | 116 | 118 | 117 | 117 | 119 | 120 | 121 |
| 0.0 | ma | mean | | | | 476 | 476 | 476 | 476 | 478 |
| | | SD | | | | 117 | 117 | 117 | 117 | 119 |
| 0.4 | mb | mean | 467 | 468 | 470 | 468 | 468 | 476 | 481 | 481 |
| | | SD | 119 | 118 | 121 | 120 | 119 | 120 | 127 | 123 |
| 0.2 | mb | mean | 471 | 472 | 474 | 472 | 472 | 476 | 478 | 479 |
| | | SD | 120 | 119 | 121 | 119 | 119 | 119 | 122 | 122 |
| 0.0 | mb | mean | | | | 475 | 476 | 476 | 476 | 477 |
| | | SD | | | | 118 | 118 | 117 | 118 | 120 |

(Target equating for math scores: mean =477; SD = 118)

Table 2. Bias and root-mean-squared difference (RMSD) of equated scores (on SAT 200-to-800 scale) produced by each equating from those produced by equipercentile equating in the full population.

| | Sampling method: | | Matched on Anchor | | | Representative | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Equating method: | | Linear | Equi% | IRT | Tucker | Freqency est. | Chained equi% | Levine | IRT |
| Old form sample | Anchor test | | | | | | | | | |
| 0.4 | va | bias | -9.1 | -9.2 | -8.3 | -9.6 | -9.0 | -4.1 | -1.7 | -1.5 |
| | | RMSD | 9.5 | 9.8 | 8.4 | 10.5 | 9.9 | 6.3 | 2.8 | 3.0 |
| 0.2 | va | bias | -3.6 | -3.5 | -3.4 | -3.8 | -3.4 | -1.3 | -0.2 | -0.8 |
| | | RMSD | 4.4 | 4.8 | 3.7 | 4.7 | 4.4 | 4.4 | 2.3 | 2.6 |
| 0.0 | va | bias | | | | -1.0 | -0.9 | -1.1 | -1.0 | -1.1 |
| | | RMSD | | | | 2.5 | 1.8 | 2.7 | 2.5 | 2.0 |
| 0.4 | vb | bias | -4.6 | -4.7 | -4.8 | -5.8 | -5.0 | +0.9 | +3.6 | +3.4 |
| | | RMSD | 5.5 | 5.7 | 5.2 | 6.8 | 5.6 | 3.4 | 5.3 | 5.8 |
| 0.2 | vb | bias | -2.7 | -2.7 | -2.2 | +0.3 | +0.5 | +2.8 | +3.7 | +3.9 |
| | | RMSD | 3.5 | 3.3 | 3.0 | 3.5 | 3.1 | 4.4 | 4.3 | 4.3 |
| 0.0 | vb | bias | | | | +2.0 | +1.8 | +1.9 | +1.9 | +2.3 |
| | | RMSD | | | | 3.0 | 2.5 | 2.8 | 3.0 | 2.8 |
| 0.4 | ma | bias | -12.9 | -12.8 | -11.6 | -11.9 | -11.1 | -4.3 | -0.5 | +0.2 |
| | | RMSD | 13.1 | 13.7 | 12.0 | 12.2 | 11.7 | 5.1 | 3.6 | 3.8 |
| 0.2 | ma | bias | -5.8 | -5.7 | -4.8 | -6.9 | -6.5 | -3.8 | -2.2 | -0.8 |
| | | RMSD | 6.2 | 6.4 | 5.3 | 7.1 | 6.9 | 4.5 | 3.8 | 3.9 |
| 0.0 | ma | bias | | | | -0.8 | -0.8 | -0.2 | -0.2 | +1.5 |
| | | RMSD | | | | 1.9 | 1.8 | 1.9 | 1.8 | 2.4 |
| 0.4 | mb | bias | -9.2 | -8.9 | -6.8 | -9.0 | -8.7 | -1.3 | +4.4 | +4.5 |
| | | RMSD | 9.4 | 9.9 | 8.2 | 9.4 | 9.7 | 6.0 | 10.4 | 7.5 |
| 0.2 | mb | bias | -5.1 | -5.0 | -2.7 | -4.5 | -4.4 | -0.6 | +1.7 | +2.9 |
| | | RMSD | 5.7 | 5.9 | 4.4 | 4.9 | 5.5 | 2.7 | 4.9 | 5.6 |
| 0.0 | mb | bias | | | | -1.3 | -1.3 | -1.0 | -1.1 | +0.7 |
| | | RMSD | | | | 2.1 | 1.9 | 1.5 | 1.9 | 3.2 |

Table 3. Bias and root-mean-squared difference (RMSD), in the full population, of equated scores (on SAT 200-to-800 scale) produced by each equating from those produced by an equating using the same method applied in the subpopulation of all students taking the anchor test.

| | | Sampling method: | Matched on Anchor | | | Representative | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Equating method: | Linear | Equi% | IRT | Tucker | Freqency est. | Chained equi% | Levine | IRT |
| Old form sample | Anchor test | | | | | | | | | |
| 0.4 | va | bias | -8.1 | -8.1 | -7.2 | -8.6 | -8.1 | -3.0 | -0.7 | -0.4 |
| | | RMSD | 8.3 | 8.8 | 7.4 | 9.5 | 9.0 | 5.1 | 0.8 | 1.4 |
| 0.2 | va | bias | -2.6 | -2.4 | -2.3 | -2.8 | -2.4 | -0.2 | +0.8 | +0.3 |
| | | RMSD | 3.1 | 4.0 | 2.6 | 3.4 | 3.4 | 2.9 | 0.9 | 1.3 |
| 0.4 | vb | bias | -6.6 | -6.7 | -7.1 | -7.7 | -6.9 | -1.1 | +1.6 | +1.1 |
| | | RMSD | 7.0 | 7.9 | 7.4 | 8.4 | 7.5 | 2.6 | 3.4 | 3.7 |
| 0.2 | vb | bias | -4.7 | -4.6 | -4.5 | -1.7 | -1.3 | +0.9 | +1.8 | +1.6 |
| | | RMSD | 4.8 | 5.0 | 4.7 | 3.5 | 3.6 | 2.6 | 1.8 | 1.7 |
| 0.4 | ma | bias | -12.2 | -12.6 | -13.1 | -11.1 | -10.3 | -4.1 | -0.4 | -1.4 |
| | | RMSD | 12.2 | 13.5 | 13.3 | 11.2 | 10.9 | 5.3 | 4.1 | 3.0 |
| 0.2 | ma | bias | -5.0 | -5.5 | -6.4 | -6.1 | -5.7 | -3.5 | -2.0 | -2.4 |
| | | RMSD | 5.1 | 6.2 | 6.5 | 6.2 | 6.2 | 4.7 | 4.1 | 3.5 |
| 0.4 | mb | bias | -7.9 | -8.0 | -7.5 | -7.7 | -7.5 | -0.3 | +5.5 | +3.8 |
| | | RMSD | 7.9 | 9.0 | 7.7 | 7.9 | 8.4 | 6.2 | 10.9 | 6.1 |
| 0.2 | mb | bias | -3.8 | -4.1 | -3.4 | -3.1 | -3.1 | +0.3 | +2.8 | +2.2 |
| | | RMSD | 4.0 | 5.1 | 3.8 | 3.2 | 4.1 | 2.7 | 5.2 | 3.3 |

## APPENDIX

Table A1.

Characteristics of equating samples:
Type of sample, number of students, mean anchor test score

| Test, anchor | New Form Sample | | | Old Form Sample | | |
|---|---|---|---|---|---|---|
| | Type | n | anchor mean | Type | n | anchor mean |
| Verbal, va | rep. | 3007 | 19.11 | 0.4 matched* | 1507 | 19.15 |
| | | | | 0.4 rep. | 2998 | 15.77 |
| | | | | 0.2 matched* | 3006 | 19.12 |
| | | | | 0.2 rep. | 2997 | 17.45 |
| | rep.** | 7959 | 19.16 | 0.0 rep.** | 8512 | 19.17 |
| Verbal, vb | rep. | 3004 | 18.22 | 0.4 matched* | 1525 | 18.36 |
| | | | | 0.4 rep. | 2998 | 14.80 |
| | | | | 0.2 matched* | 3004 | 18.21 |
| | | | | 0.2 rep. | 2999 | 16.69 |
| | rep.** | 7625 | 18.21 | 0.0 rep.** | 8329 | 18.23 |
| Mach, ma | rep. | 2999 | 11.36 | 0.4 matched* | 1498 | 11.36 |
| | | | | 0.4 rep. | 3002 | 8.96 |
| | | | | 0.2 matched* | 2999 | 11.36 |
| | | | | 0.2 rep. | 3000 | 10.29 |
| | rep.** | 8450 | 11,45 | 0.0 rep.** | 8000 | 11.28 |

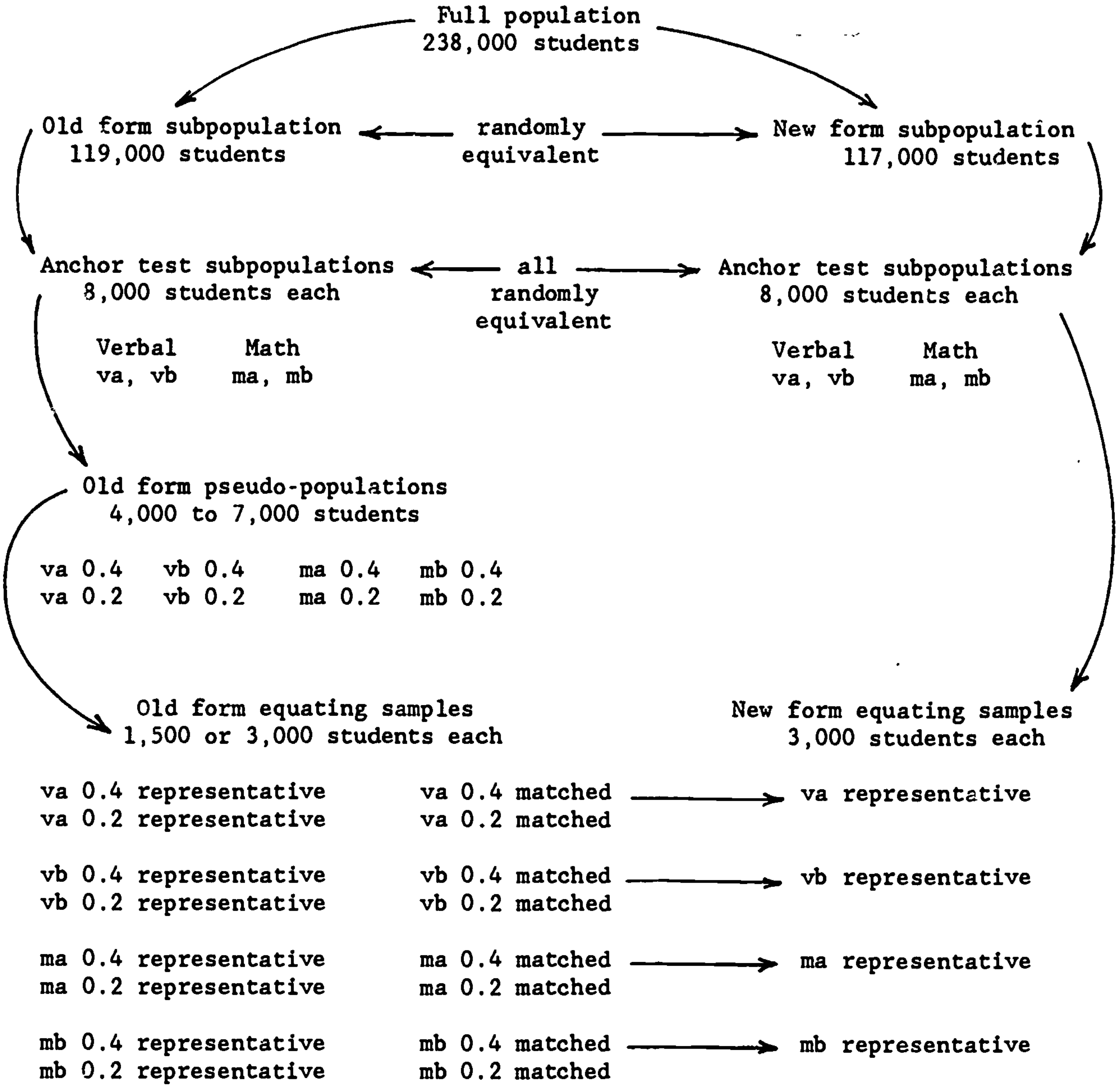
(Continued on next page)

Table Al. (continued)

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Math, mb | rep. | 2998 | 10.26 | 0.4 | matched* | 1480 | 10.20 |
| | | | | 0.4 | rep. | 3003 | 7.69 |
| | | | | 0.2 | matched* | 2998 | 10.26 |
| | | | | 0.2 | rep. | 2999 | 8.91 |
| | rep.** | 8161 | 10.20 | 0.0 | rep.** | 7764 | 10.12 |

* Imperfectly matched
** Subpopulation created by spiraling of test forms

Figure 1.' Design of the Study



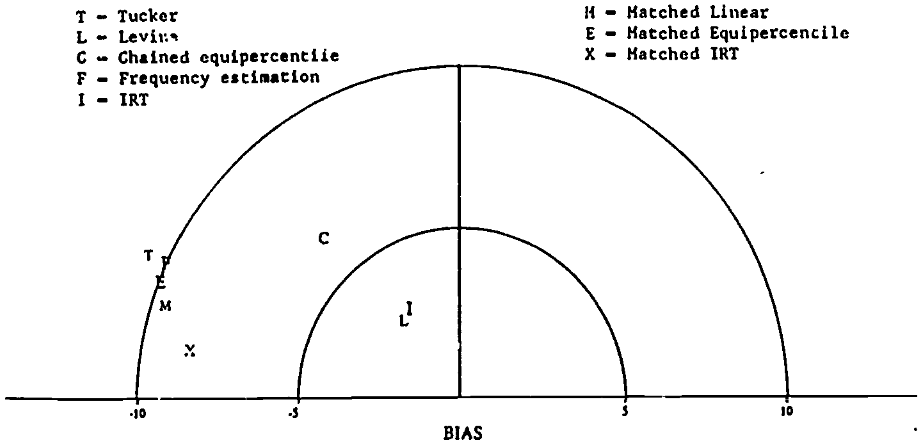


Figure 2a. Bias and RMSD in equating the verbal scores through anchor "va", sampling from the "0.4 population".

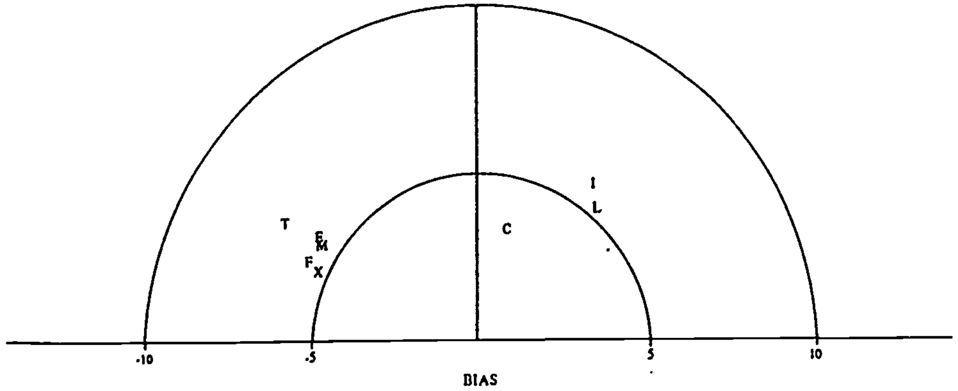


Figure 2b. Bias and RMSO in equating the verbal scores through anchor "vb", sampling from the "0.4 population".

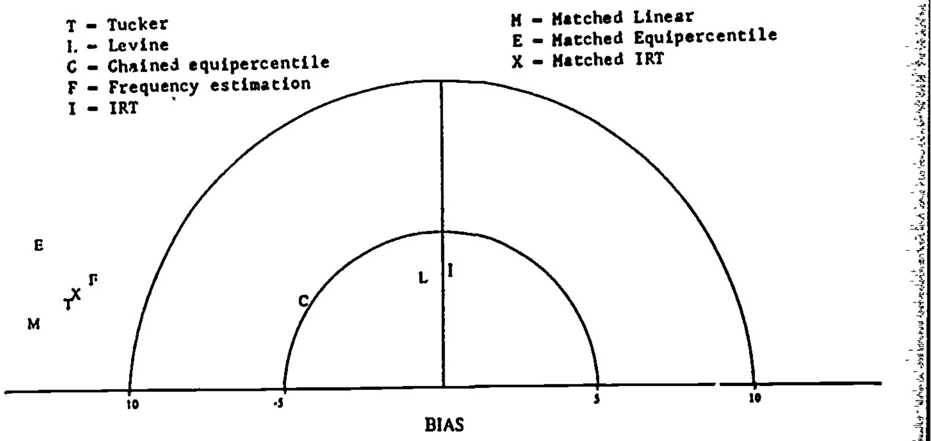


Figure 2c. Bias and RMSD in equating the math scores through anchor "ma", sampling from the "0.4 population".

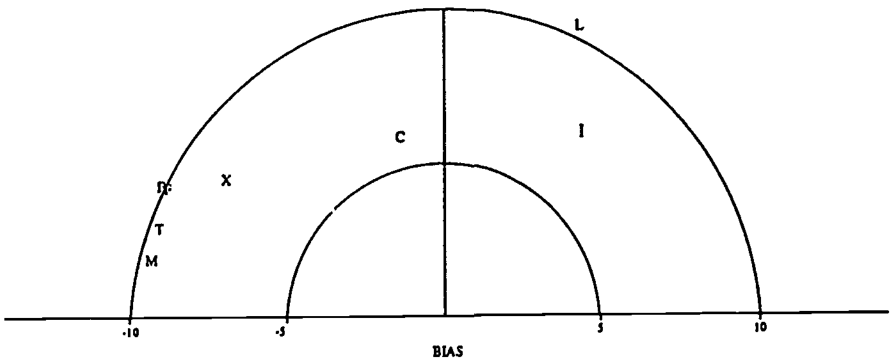


Figure 2d. Bias and RMSO in equating the math scores through anchor "mb", sampling from the "0.4 population".

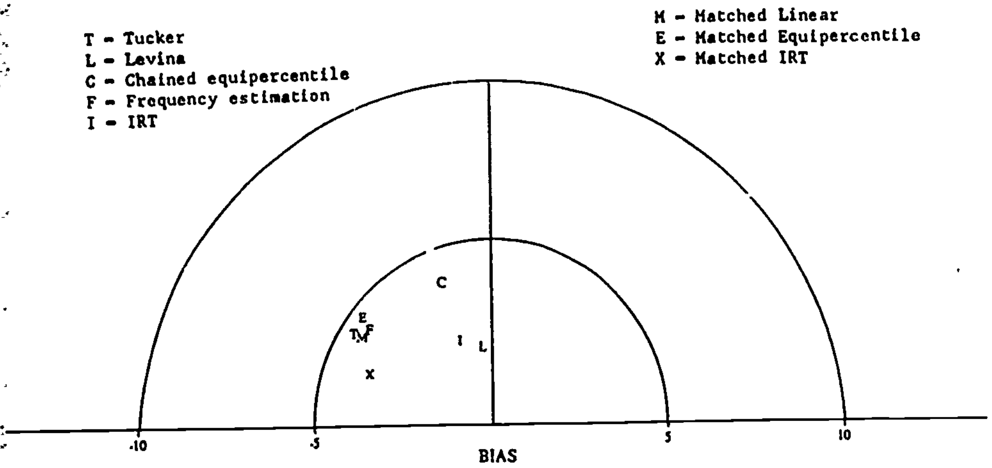


Figure 3a. Bias and RMSD in equating the verbal scores through anchor "va", sampling from the "0.2 population".

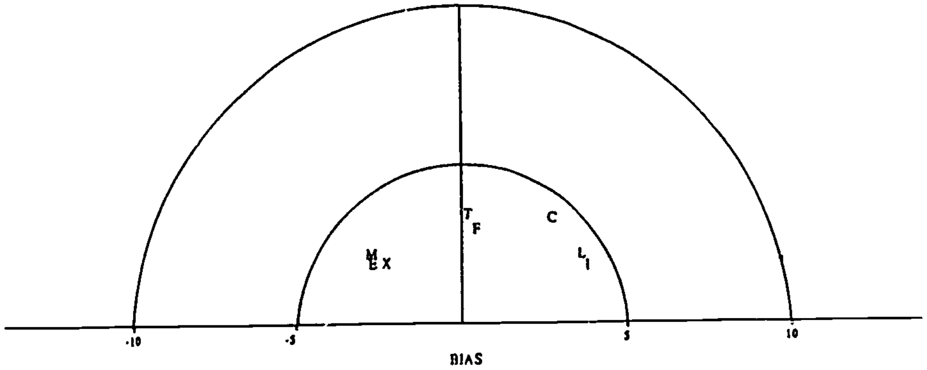


Figure 3b. Bias and RMSD in equating the verbal scores through anchor "vb", sampling from the "0.2 population".

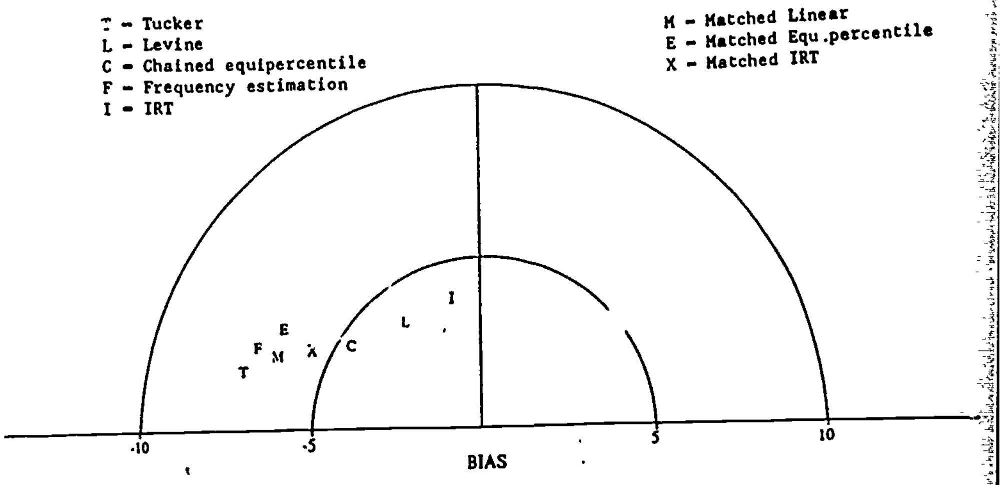


Figure 3c. Bias and RMSD in equating the math scores through anchor "ma", sampling from the "0.2 population".

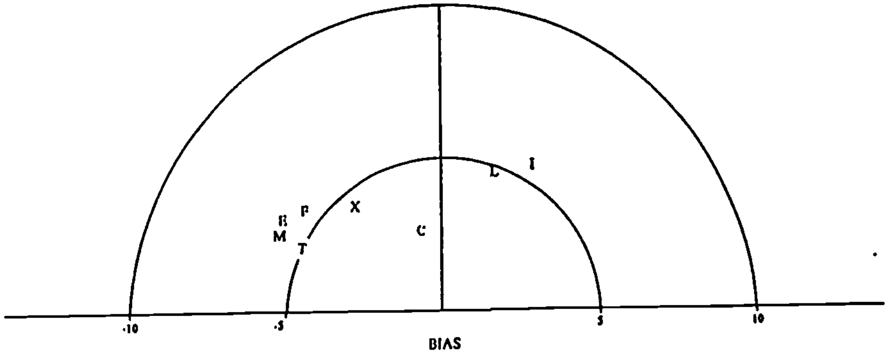


Figure 3d. Bias and RMSD in equating the math scores through anchor "mb", sampling from the "0.2 population".

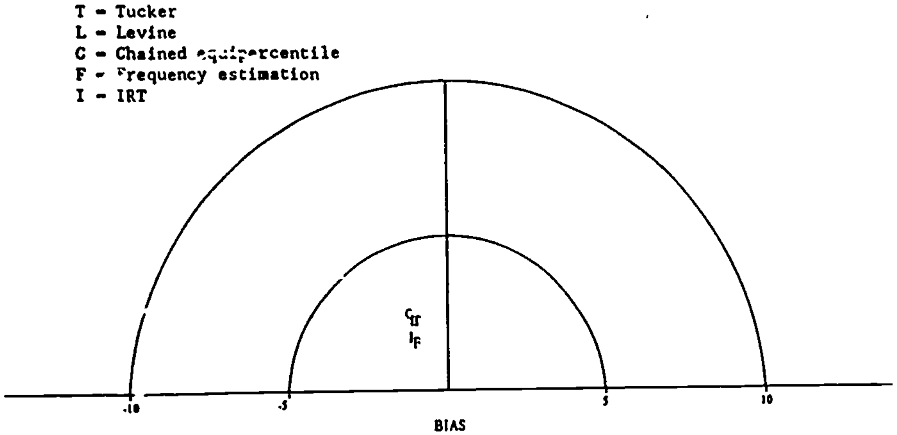


Figure 4a. Bias and RMSD in equating the verbal scores th.ough anchor "va" in the anchor test subpopulations.

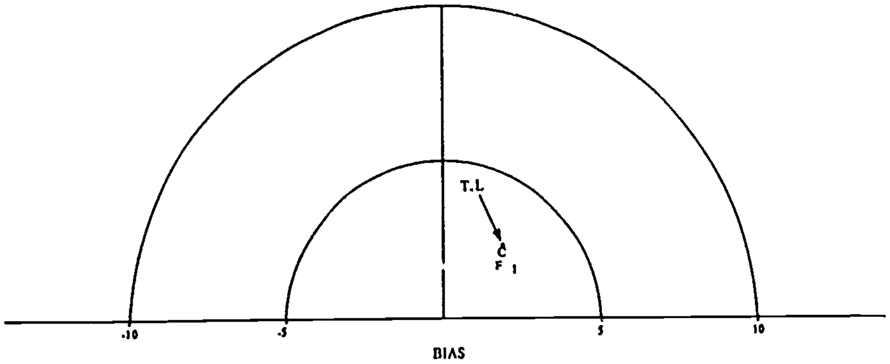


Figure 4b. Bias and RMSD in equating the verbal scores through anchor "vb" in the anchor test subpopulations.

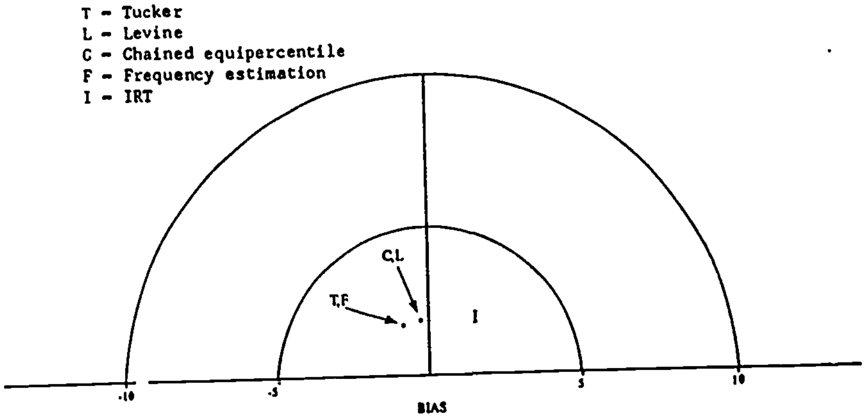


Figure 4c. Bias and RMSD in equating the math scores through anchor "ma" in the anchor test subpopulations.

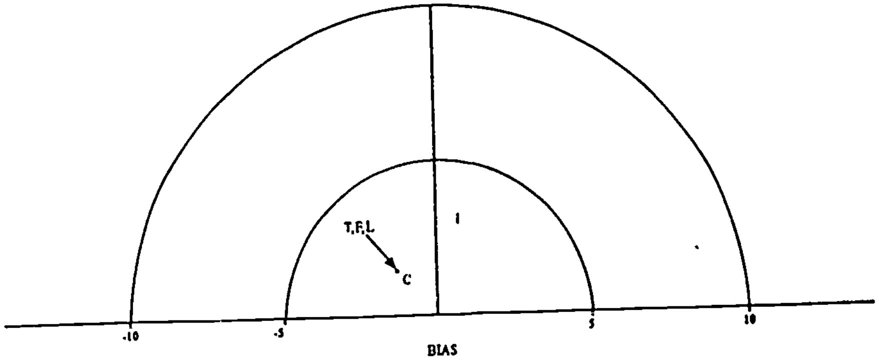


Figure 4d. Bias and RMSD in equating the math scores through anchor "mb" in the anchor test subpopulations.