ED 385 547 TM 023 963

AUTHOR	Wright, Nancy K.; Dorans, Neil J.
TITLE	Using the Selection Variable for Matching or Equating.
INSTITUTION	Educational Testing Service, Princeton, N.J.
REPORT NO	ETS-RR-93-4
PUB DATE	Jan 93
NOTE	40p.; Based on a paper presented at the Annual Meeting of the National Council on Measurement in Education (Boston, MA, April 17-19, 1990).
PUB TYPE	Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	Criteria; *Equated Scores; *Selection; Simulation; *Test Results
IDENTIFIERS	Accuracy; Anchor Tests; Chained Equipercentile Equating; Frequency Estimation Equipercentile Equating; Levine Equating Method; Scholastic Aptitude Test; Tucker Common Item Equating Method; *Variables; Variance (Statistical)

ABSTRACT

This paper studies whether equating results can be improved if the variable that accounts for all systematic differences between equating populations is identified and used as an anchor in anchor test design or as a variable on which to match equating samples. The sample invariant properties of four anchor test equating methods (Tucker and Levine equally reliable linear models, chained equipercentile, and frequency estimation equipercentile models) were examined under representative, matched-on-equating-test, and matched-on-selection-variable conditions. The selection variable, the variable along which subpopulations differ, was also used as an anchor for the four equating methods and compared to equatings in which the equating test served as the anchor. All equatings were performed with real Scholastic Aptitude Test (SAT) populations or simulated populations. Accuracy criteria were equivalent-groups equipercentile equatings based on old and new form subpopulations of over 115,000 test takers. Results showed that matching on the selection variable improved accuracy over matching on the equating test for all methods. Results with the selection variable as an anchor were good for both the Tucker and frequency estimation methods, but unacceptable for chained equipercentile and Levine results. Two tables, 2 figures, and 16 graphs present analysis data. (Contains 24 references.) (Author/SLD)

RR-93-4

RESEARCH REPORT

USING THE SELECTION VARIABLE FOR MATCHING OR EQUATING

Nancy K. Wright
Neil J. Dorans

Educational Testing Service
Princeton, New Jersey
January 1993

Using the Selection Variable for Matching or Equating [1,2]

Nancy K. Wright and Neil J. Dorans
Educational Testing Service

December 1992

[1] This report is based on a paper presented at the 1990 annual meeting of the National Council on Measurement in Education, Boston.

[2] The authors thank Samuel Livingston for his review of an earlier version of this paper, and Elizabeth Burton, Miriam Feigenbaum and Bernadette McIntosh for their assistance with data analyses, and Nancy S. Petersen and Neal Thomas for their review of the preceeding version of this manuscript. The opinions expressed are those of the authors, and do not represent official positions of the Educational Testing Service. Funding for this paper was provided by the Admissions Testing Program of the Educational Testing Service.

Abstract

This paper addresses whether we can improve equating results if we know the variable that accounts for all systematic differences between equating populations and use it as either an anchor in an anchor test design or as a variable on which to match equating samples. The sample invariant properties of four anchor test equating methods (Tucker and Levine equally reliable linear models, chained equipercentile and frequency estimation equipercentile models) under three sampling conditions, "representative", "matched on equating test", and "matched on selection variable" are examined. The "selection variable" is defined as the variable or set of variables along which subpopulations differ. In addition to being used for matching of subpopulations, the selection variable was used as an anchor for the four equating methods, and compared to equatings in which the equating test served as the anchor. All equatings were performed with either real Scholastic Aptitude Test (SAT) populations or with data drawn from simulated pseudopopulations which differed from their original real SAT populations on the basis of the selection variable. The tests used were the verbal and math portions of two forms of the SAT. The criteria for accuracy were equivalent-groups equipercentile equatings based on old and new form subpopulations of over 115,000 test takers.

Results showed that matching on the selection variable improved accuracy over matching on the equating test for all methods. Compared with the representative sample equatings, Tucker and frequency estimation results improved with matching on the selection variable; chained equipercentile and Levine results were similar under these two sampling conditions. Results with the selection variable as an anchor were good for both the Tucker and frequency estimation methods; chained equipercentile and Levine results were quite unacceptable as anticipated since use of the selection variable -- math scores for the verbal equatings and verbal scores for the math equatings -- violated assumptions of these models.

The positive results obtained for use of the selection variable as a matching variable or anchor test (for some methods) suggest that future research into the reasons test takers select certain test administrations may lead to improved test score equating practices.

Using the Selection Variable for Matching or Equating

Nancy K. Wright and Neil J. Dorans
Educational Testing Service

Recent studies on the sensitivity of equating methods to sampling and subpopulation differences suggest that no currently used equating method that employs an anchor test design is always able to achieve sample and subpopulation invariant equating results. When samples from different subpopulations vary along a dimension that differs from that measured by the equating test -- true ability or another set of variables -- some equating methods do not adjust properly for sample differences and hence differences in test difficulty. This paper is a follow-up investigation to a set of papers that examined the efficacy of using equating test scores to match equating samples in order to improve the performance of different equating methods that use an anchor test design in which the equating samples are quite dissimilar in ability. The present study uses data from one of these papers, the Livingston, Dorans and Wright (1990) simulation study, where the selection variable is known, to assess whether matching with or equating with the selection variable can be used to produce acceptably invariant equatings.

First, relevant previous research is reviewed in order to set the stage for the present investigation. Then, the purpose of the present study is described. Next, the equating methods employed in the present study are described. The design of the present study is then presented with particular emphasis placed on what it adds to the previous research. Results are presented, and finally discussed.

Background and Purpose

Background

Interest in score equating procedures, both from a technical and practical point of view, has been increasing steadily in recent years. Brennan (1987), in the lead article of a 1987 issue of Applied Psychological Measurement, a substantial portion of which was devoted to *Problems, Perspectives, and Practical Issues in Equating*, gave two reasons for this surge of interest. First, there has been an increase in the number of testing programs that use multiple editions of the same test, prompting an increased awareness that equating is needed to ensure that scores are equitable across different editions of the test. Second, test publishers and developers have had to reference the role of score equating in a climate of enhanced public scrutiny over standardized testing. In addition, research in equating has focused on improved equating methods, particularly, item response theory (IRT) methods. Much of this IRT research is summarized in Skaggs and Lissitz (1986). The Cook and Petersen (1987) article in the aforementioned AME special issue also included a review of studies about several different equating methods including IRT. References to more recent equating research can be found through the scaling, norming and equating chapter by Petersen, Kolen, and Hoover (1989) in the most recent edition of Educational Measurement (Linn, 1989). Most of this earlier research examined how well equating methods performed with intact representative samples.

Matching of equating samples initially was viewed as a potential solution to the chronic problem of diverging equating results obtained under different equating methods when data are collected in an anchor test design in which the old form and new form equating samples differ substantially (Lawrence & Dorans, 1988). At the 1989 annual meeting of the AERA, a symposium, entitled *Selecting Samples for Equating: To Match or Not to Match,* focused on how different equating methods perform under different sampling conditions: representative or matched samples. That symposium evolved into a special issue of *Applied Measurement in Education* (Dorans, 1990a). Since the present study is a follow-up to research in that symposium / special issue, those studies will be summarized here.

In the lead article, Dorans (1990b) described the equating methods used and sampling designs employed by the empirical studies in that special issue. Four requisites for equating were listed and the invariance of equating functions requisite was identified as the focus of the special issue. Descriptions were given of the Tucker equating method, the Levine equally reliable and unequally reliable equating methods, the chained equipercentile equating method, the frequency estimation equipercentile and linear equating methods, and the three parameter logistic (3PL) item response theory true-score equating method. All these methods employ data collected within an anchor test design. The description of each method focused on assumptions made by that method, included some basic mathematical expressions associated with the method, and described procedural aspects of the method. Similarities and differences among methods were also discussed. Three types of sampling designs were described: representative sampling; new-form matched sampling (old form sample to new form sample); and reference or target matched sampling (old and new samples to a reference population). Some of the practical mechanics of matching were discussed.

Using data from several administrations of the Scholastic Aptitude Test (SAT), Lawrence and Dorans (1990) addressed the sample invariant properties of five anchor test equating methods across two sampling conditions to see which methods 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 new form 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 method differed for representative and matched samples, as did results for the Levine equally reliable and chained equipercentile methods. Results based on the Tucker observed-score method and frequency estimation equipercentile equating method were found to be essentially invariant across representative and new form matched sample conditions. Results for the five equating methods tended to converge under the new form matched sample condition. Tentative explanations for the findings were offered.

Eignor, Stocking and Cook (1990) employed a simulation model to study the invariance effect. Two independent replications of a sequence of simulations were carried out to evaluate the performance of four anchor test equating methods under two sampling design conditions. Since the data were generated according to an item response theory model, it was predicted that the IRT equating method and the true-score Levine equally reliable equating method would be less affected by sample differences, and the results confirmed this finding. The authors advised against matching on equating tests for the IRT, Levine equally reliable and chained equipercentile methods.

Schmitt, Cook, Dorans and Eignor (1990) examined the results of equating two parallel editions of an Achievement Test in Biology using different equating methods under different sampling strategies. In addition to representative samples and new form matched samples, they studied reference or target matched sampling. The criterion equating was a Tucker equating using representative samples from two populations that were very close in ability. They found that matching on a set of common items provided greater agreement among the results of the various equating procedures than was obtained under representative sampling. In addition, for all equating procedures, the results of equating with samples matched on common item scores agreed more closely with the criterion equating than did results from representative samples. Matching to an reference target population produced agreement among methods, but did not agree as closely with the criterion equating as matching to the new form on the basis of common item scores. The equating models least affected by differences in new and old form sample abilities were the Tucker and frequency estimation equipercentile models, and the procedure most affected by ability differences was the IRT procedure. (Cook, Eignor & Schmitt (1989) examined one edition of four other Achievement Tests and failed to replicate the superiority of matched sample equatings.)

Livingston, Dorans and Wright (1990) examined five equating methods under two sampling conditions using data specially constructed from a national administration of the SAT. The criterion equating was based on an equivalent-groups design equating involving more than 115,000 students taking each of two editions of the SAT. 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 equally reliable methods tended to agree with each other and were inconsistent in the direction of their bias.

This set of papers could be viewed as a psychometric drama about the efficacy of matching, which swayed from a "yes" based on the Lawrence and Dorans (1990) study to a definite "no" according to Eignor, Stocking and Cook (1990), back to a "yes" by Schmitt, Cook, Dorans and Eignor (1990), then back yet again to "no" according to Livingston, Dorans and Wright (1990). Kolen (1990) and Skaggs (1990) examined these articles, synthesized them, posed questions, and discussed their implications for current and future equating practices. In addition to providing critiques of the individual articles, both Kolen and Skaggs looked for universal themes that could be extracted from this psychometric drama.

Skaggs (1990) concluded that Tucker and frequency estimation are not affected by matching on the equating test, while Levine, IRT and chained equipercentile equating are affected. Skaggs also pointed out that the conclusions one might draw about the efficacy of matching depend on the criterion used. If consistency among methods is the criterion, then matching achieves that consistency. Skaggs also raises the issue of multidimensionality and wonders how it may have affected different methods in the different studies. Finally, he concludes that we need to know more about examinees and how they end up in samples. Until we know that, "...matching appears to be a risky business."

Kolen (1990) indicated that there were three general research findings that underlie this set of studies. First, when equivalent groups of examinees are given carefully constructed test forms, the equating relationship is invariant with respect to equating populations. Second, when an anchor test is used in which the anchor is a miniature of the total test form and is administered to groups taking the old and new form who are similar to each other, then equating methods tend to give similar results. Third, when an anchor test is used and the groups taking the old and new forms are quite different, then any equating method may give poor results. A major motivation for these related studies was an attempt to improve anchor test equatings when the old form and new form groups were quite dissimilar via matching on the equating test. Kolen concludes from these studies that matching on the equating test does not result in more accurate equating. He also states that matching on other variables is worthy of future research.

Purpose

This study attempts to begin to address the issue of whether matching on something other than the equating test can produce more accurate equating results. Livingston, Dorans and Wright (1990) offered an avenue for future research: if matching on the anchor score is not a good idea, a promising variable on which to match equating samples might be the selection variable or set of variables that causes the new and old form groups to differ systematically. In practice, we don't know the selection variable, but we might be able to model it. A "propensity score" (Rosenbaum & Rubin, 1985), a linear combination of all the variables we can measure that best discriminates between the two populations, may be a promising way of modelling the self-selection process. Collecting, constructing and matching on propensity scores is a more complicated procedure than matching on anchor scores alone. As a preliminary step before studying the use of propensity scores in the equating process, this study examines whether a **known** selection variable, such as the variable actually used to create populations in the Livingston, Dorans and Wright (1990) study, can produce acceptable equating results if used as a matching variable or as an anchor score distribution in the equating process. It assesses whether knowledge of the process underlying the self-selection of students to administrations can be used to improve equating results with certain equating models. If knowledge of the self-selection process can be used to improve equating results, then we can focus future research efforts on attempting to model self-selection.

The Equating Methods Used in Our Anchor Test Design

Anchor Test Design

The old form equating sample, which takes the old form (X), and the new form equating sample, which takes the new form (Y), can be related to each other in one of three ways: (1) the old form sample and new form sample are identical, the "single-group" design; (2) the old form and new form samples are statistically exchangeable, the "equivalent-groups" design; (3) the old and new form samples are not statistically exchangeable, the "non-equivalent-groups" design (Angoff, 1984; Petersen, Kolen & Hoover, 1989).

In the non-exchangeable-groups design or anchor test design, one group takes the old form and another group takes the new form, but the samples are not selected to ensure equivalent test performance. Ordinarily, the equating data come from different test administrations. Equating tests or anchor tests are essential for designs in which the old form and new form samples are not exchangeable. This paper uses this third data collection design.

Equating Methods

In this study, four equating methods were employed: chained equipercentile equating, frequency estimation equipercentile equating, Tucker linear equating and Levine equally reliable equating.

Equipercentile equating methods. The equipercentile equating function, e(y), equates test Y to test X on some population P if test X and e(y) have the same cumulative frequency distribution on population P. For obvious reasons, equipercentile equating is also referred to as distribution matching. Equipercentile equating is based on the definition that the score scales for two tests are comparable if the score distributions for the two tests are identical in shape for some population P (Braun & Holland, 1982).

Equipercentile equating can be viewed as a two-stage process (Kolen, 1984). First, the relative cumulative frequency distributions are tabulated or plotted for the two forms to be equated. Second, equated scores are obtained from these relative cumulative frequency distributions. A cumulative distribution function maps scores onto relative frequencies which have a maximum of 1 and a minimum of 0. An inverse cumulative distribution function maps frequencies onto scores.

One equipercentile method that uses an equating test is what Angoff (1984, pg. 116) refers to as design V, what Braun and Holland (1982, pp. 39-42) call equating two tests through a third test, and what we refer to as the chained equipercentile method. In subpopulation P, test Y is equated to equating test V such that equated scores refer to the same percentile rank of examinees in P. In subpopulation Q, test X is equated to equating test V such that equated scores refer to the same percentile rank of examinees in Q. Scores on X and Y are said to be equated if they correspond to the same score on anchor test V. Note that two separate equatings are actually employed in two different subpopulations and that test X and Y are never directly equated in a single population. For this method to make sense from an equating point of view, one must assume that the new form sample and the old form sample are both representative of a common population, i. e., P and Q are identical. In practice, however, this method is used even when P and Q are not identical.

Another equipercentile equating procedure that uses an anchor test is called frequency estimation (Angoff, 1984 p. 113). This procedure attempts to simulate a situation in which both X and Y are taken by a single or exchangeable groups. Data from P and Q on V are combined and used to estimate the frequencies on X and Y that this combined group would have obtained had they taken both X and Y. 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 synthetic population, R, with a specified distribution of scores on the anchor test, typically the distribution in the combined (old form and new form) sample. The key assumption is that the conditional distribution of scores on the new form (Y), given the score on the anchor test (V), 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 (X). 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. Once these frequencies have been estimated, a standard equipercentile equating of Y to X is performed to obtain e(y) on this combined population.

Linear equating methods. Linear equating can be viewed as a very smoothed version of equipercentile equating in which only the first two moments of the score distributions of X and Y on P are matched. In linear equating, a transformation is found such that scores on X and Y are said to be equated if they correspond to the same number of standard deviation units above and below the mean in P.

There are a variety of linear equating models that employ an equating test. The volume edited by Holland and Rubin (1982) contains several chapters that describe these various models; in particular, chapters by Angoff (1982), Petersen, Marco and Stewart (1982), and Potthoff (1982) should be consulted. Two of the more popular models are the Tucker model and the Levine equally reliable model.

The Tucker linear equating model assumes that the regression of total score Y onto the equating test V is linear and homoscedastic, and that this regression, which is observed in the sample that took test Y with V, also holds in the sample that took test X with V. A similar set of assumptions is made about the regression of X on V.

The Levine equally reliable linear equating model assumes that the true scores on Y and V are perfectly related, and that the ratio of the standard deviation of true scores on Y to the standard deviation of true scores on V is the same in the observed group P and the synthetic population R. In addition, it assumes that the intercept of the regression line relating true scores on Y to true scores on V is the same in P and R. Further it assumes that the standard error of measurement for Y and for V is the same for groups P and R. A similar set of assumptions are made about true scores on X and V in the observed group Q and R. A common misconception holds that the Levine equally reliable equating method is a true score equating method. It is not. It estimates observed score means and standard deviations using assumptions about true score regressions and standard errors of measurement. Hence, it is an observed score equating method based on assumptions about true scores.

Design of Study

Data Source

The tests used in this study and in Livingston, Dorans and Wright (1990) were two forms of the verbal and mathematical portions of the SAT administered concurrently at a large national administration by alternating or "spiraling" the two forms. The population of test takers for both studies was restricted to high school juniors and seniors, the target population for the SAT and the one used in the operational equatings of the tests. A total of approximately 236,000 juniors and seniors took the forms: 119,000 examinees took one form and 117,000 the other. The equipercentile relationship in the raw score distributions on these two forms for populations this large could be expected to represent the true equating relationship as nearly as possible. The raw-to-scale version of this equipercentile equating was used as the criterion against which the various experimental equatings were evaluated for accuracy.

For the purposes of the studies one of the forms was assigned to be the "new form" and the other the "old form." No anchor was needed to equate these forms since groups administered two forms by alternating booklets could be considered random groups from the same population. However, four equating tests in the form of external common-item sets -- two verbal and two math -- were administered to random subgroups on each of the two forms for the purpose of equating the forms to past and future editions of the SAT. The common-ıtem equating sets were each parallel in content and length to one of the operational sections of the verbal or math operational tests. The equating tests were systematically spiraled in test booklets to form eight stratified random subgroups of approximately 8,000 students each. Each subgroup was smaller than one-eighth of the population because pretests rather than equating tests were administered to some students.

Generation of Pseudopopulations

Data simulated for the Livingston, Dorans and Wright (1990) study were utilized in the present study. The following describes the data simulation:

> ....The four anchor tests made it possible to create, artificially, several anchor equating situatior in which the populations of students taking the old form differed systematically in ability from the populations taking the new form. Most important, in each of these anchor equating situations the true equating relationship in the target population was known (or rather, to be strictly correct, this relationship could be very precisely estimated). Each equating situation consisted of a pair of populations linked by an anchor test. The new-form population in each pair was simply the subpopulation of students taking the new form and the anchor test. Each old-form population was actually a pseudopopulation selected to be of systematically lower ability than the new form population. The old-form pseudopopulation in each pair was selected from the subpopulation of students taking the old form and the anchor test, by removing a portion of the higher ability students. The old-form pseudopopulations for equating the Verbal test were selected on the basis of their Math scores, to avoid selecting on either the anchor (equating) score or the score to be equated. Similarly, the old-form pseudopopulations for equating the Math test were selected on the basis of their Verbal scores.
>
> Each new-form population was paired with two different old-form psedopopulations of different ability levels. One of the old-form psuedopopulations was selected to have a mean ability level approximately 0.2 SD lower than the new-form psuedopopulation. This psuedopopulation is referred to as the *0.2 population*. The other old-form pseudopopulation was selected to have a mean ability level approximately

> 0.4 SD lower than the new-form population. This old-form pseudopopulation is referred to as the *0.4 population.*[1] The 0.2 populations varied in size from 6,148 to 6,658 students; the 0.4 populations varied in size from 4,367 to 4,887. (p. 76-78).

The .2 SD condition is often seen in practice with SAT data, while the .4 SD condition is seen on occasion.

Samples for Equating

The four new form samples for all experimental equatings in this study andin Livingston, Dorans and Wright (1990) consisted of approximately 3,000 test takers each, selected by a technique called "spaced random sampling" from the full new form equating test subpopulations. Spaced random sampling involves dividing the full group into blocks or "spaces" of equal size and randomly selecting an equal number from each block so that the desired sample size is obtained. Eight old form samples were selected in like manner, four from the "0.2 populations" and four from the "0.4 populations." These twelve samples -- eight old and four new - - will be referred to as representative samples.

In addition, two sets of matched samples were selected. The first set, selected for and used in Livingston, Dorans and Wright (1990) consisted of four old form samples of approximately 3,000 test takers each from the "0.2 populations," each matched to the appropriate new form representative sample using the common-item equating test as a stratifying variable. It was not possible to select perfectly matched samples in any of these four cases; that is, the number in each sample and the equating test mean for each old form group varied slightly from the corresponding new form sample.

A second set of four matched old form samples from the "0.2 populations" was selected for the present study using the selection variable distribution -- the variable used in the simulation of the pseudopopulations -- in each new form sample as the stratifying variable. Math raw scores which had been transformed to

[1] 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 SD below that of the full population. A similar procedure was used for selecting the other old-form pseudopopulations.

the College Board 200-to-800 scale were used to select verbal matched old form samples; similarly, verbal scaled scores were used in the selection of math matched samples. In the process of creating the 0.2 pseudopopulations too many high-ability test takers had been removed to allow for fully matched samples of 3,000. Instead, proportional matching was performed: approximately, two-thirds of the cases at each score level on the stratifying variable were selected for a total sample size of approximately 2,000.

For Livingston, Dorans and Wright (1990) matched old form samples were selected from the "0.4 populations," matching on the new form equating test distribution. For the present study an attempt was made to select proportional samples from the "0.4 populations" matched on the new form selection variable distributions. However, it would have been necessary to decrease sample sizes to 750, too small a number for stable equating results in this situation.

The eight randomly occurring equating test subpopulations of approximately 8,000 test takers each -- four old form and four new form groups -- also served as equating samples. These subpopulations are labeled "0.0 populations" in tables and figures.

Figure 1 presents relationships among populations and samples used in the Livingston, Dorans and Wright (1990) study. Additional samples selected for this follow-up study are shown in **boldface**.

Insert Figure 1 about here

Choice of Anchor

The anchor test used in the equatings in Livingston, Dorans and Wright (1990) and the present follow-up study are as follows:

Livingston, Dorans and Wright (1990):	Anchor Test
Representative samples (0.0, 0.2 and 0.4 populations)	Equating test
Samples matched on equating test (0.2 and 0.4 populations)	Equating test
Present study:	Anchor Test
Representative samples (0.0, 0.2 and 0.4 populations)	Selection variable
Samples matched on selection variable (0.2 population)	Equating test

The anchor test used to link the old and new form samples in all equatings performed for the Livingston, Dorans and Wright (1990) study was the external common-item set -- the equating test. Newly performed equatings for the present study used the equating test as the anchor only in the equatings involving the four old form samples from the "0.2 populations" which had been matched to the new form samples using the selection variable score distributions.

For the equating situations in the present study using representative samples from the old form "0.0, 0.2 and 0.4 populations", the anchor test was the selection variable -- operational math scores for the verbal equatings and operational verbal scores for the math equatings. Raw scores could not be used as an anchor in these equatings because the old and new form groups had taken different operational tests. Instead, scores expressed on the College Board 200-to-800 scale were used. These scores functioned as a common anchor test because, once raw scores have been transformed to a common scale, the resulting scaled scores at a given score level (e.g. 500) on two forms can be considered interchangeable.

Criteria for Accuracy

Two criteria for evaluating equating results are described in this section. The present study uses the same approach for evaluating equatings as Livingston, Dorans and Wright (1990). The primary criterion for judging the overall accuracy of each equating was the root mean-weighted square difference (RMWSD) of the equated scores for the full new form population from their equated scores determined by the target equating. The RMWSD is computed by the formula:

$$\mathrm{RMWSD} = [(\Sigma \mathrm{n}(\mathrm{y})\{(x(y)\text{-}\mathrm{X}(\mathrm{y}))\}^{2})/\Sigma \mathrm{n}(\mathrm{y})]^{.5},$$

where n(y) is the number of examinees with raw score y on the new form, X(y) is the corresponding exact (unrounded) scaled score on the old form as determined by the target equating, and *x(y)* is the corresponding exact (unrounded) scaled score on the old form as determined by the other equating to be compared with the target 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 RMWSD statistics are in terms of this scale.

A secondary criterion for evaluating the accuracy of each equating was its bias. Bias may be described as the tendency for the equated scores to be systematically too high or too low. The overall bias statistic is an average value for the new form population. The bias of an equating is computed by the formula:

$$\text{Bias} = (\Sigma n(y)\{(x(y)\text{-}X(y)\})/\Sigma n(y)\ ,$$

where the symbols have the same meaning as in the formula for the RMWSD. Negative bias in one part of the score range may cancel out positive bias in another part of the score range. Values of the RMWSD and bias statistics of five or more are considered to indicate problematic equating results.

Results

Emphasis in this section is on the relative accuracy of equatings within method, comparing results for the present study with results from Livingston, Dorans and Wright (1990). First, relative accuracy will be presented for equatings in which the common-item equating test, used in the former study as the anchor test, has been replaced with the scaled selection variable. Next, results in samples matched on the selection variable are compared with the accuracy within method of results under two sampling conditions, matched on the equating test and representative, studied in Livingston, Dorans and Wright (1990). The reader is referred to Livingston, Dorans and Wright (1990) for a full discussion of the accuracy of various combinations of five equating methods and the latter two sampling conditions.

Equating Through the Selection Variable

The right side of Table 1 shows the bias and RMWSD statistics expressed on the College Board 200-to-800 scale for each of the equatings performed for the present study in representative samples using the selection variable as anchor. Contrasted on the left side of Table 1 are bias and RMWSD values for equating methods in which the equating test was used as anchor. Information on the left is from the Livingston, Dorans and Wright (1990) study. All indices were calculated in the full new form population, using the random-groups equipercentile equating in the full population as the criterion.

Insert Table 1 about here

Four equating methods are presented from the former study: Tucker, Levine, chained equipercentile and frequency estimation equipercentile. Only two equating methods, Tucker and frequency estimation, are shown for the equatings through the selection variable for the present study. Levine and chained equipercentile equatings were also performed. For these two methods, large values for bias and

RMWSD resulted -- some in excess of 50 scaled score points. Figure 2 shows the conversion results in the "mb 0.4 population" for all four equating methods, using the selection variable as anchor, compared with the criterion equating. While the Tucker and frequency estimation conversions follow the criterion closely, Levine and chained equipercentile diverge from the criterion in the top half of the scale by as much as 80 and 100 points, respectively -- clearly unacceptable results.

Insert Figure 2 about here

The values in Table 1 are presented graphically in Figures 3a to 5d. Figures 3a to 3d compare the accuracy of four equating methods using the common-item equating test with two methods using the selection variable, sampling in the old form "0.4 populations." Each plot presents the results for one old form sample, with the set of four representing replications. Each of the six bars in a plot shows the accuracy of a particular combination of equating method and anchor test type, as indicated by the RMWSD statistic -- the height of the bar. The shaded solid portion of the bar is the overall bias in the equating, which is always less than the RMWSD. Black indicates negative bias, while gray shading indicates positive bias. In the Livingston, Dorans and Wright (1990) study, Tucker and frequency estimation equating results tended to cluster and to exhibit unacceptably large negative bias in the 0.4 samples. In three of the four samples in the current study, these two methods also tended to be similar in accuracy. In all four replications the accuracy of Tucker results was improved, in some cases rather dramatically, when the selection variable replaced the equating test. For the frequency estimation method, accuracy was improved in three of the four cases. In the fourth, RMWSD values were about equivalent, but the overall bias was slightly smaller. In the present study, Tucker and frequency estimation results are about as accurate, in general, as chained equipercentile and Levine results from the former study.

Insert Figure 3 about here

Figures 4a to 4d present results for the same equating methods and anchor test types in the "0.2 populations." In general, more accurate equatings across method and anchor test can be observed in these samples, but there are some exceptions. As in the "0.4" equatings, Tucker and frequency estimation results tend to cluster when either the equating test or the selection variable serves as anchor.

The improvement within method of using the selection variable is both less dramatic and less clear within these samples. Only two of four equatings were more accurate for both Tucker and frequency estimation with use of the selection variable. However, in three of four cases, the new Tucker and frequency estimation were within the 5-point accuracy band and clustered with the Levine results. In the mb sample, Tucker and frequency estimation through the selection variable were less accurate than any of the methods using the equating test.

Insert Figure 4 about here

Equating results for the full randomly equivalent subpopulations are displayed in figures 5a to 5d. All equatings performed in these groups are acceptably accurate regardless of method or anchor test. In other words, when the samples in which the equatings are to be performed are close in ability, little adjustment is to be made, and most methods work well regardless of anchor used. Differences observed in these equatings are likely due to sampling variability.

Insert Figure 5 about here

Matching Through the Selection Variable

Table 2 displays accuracy values for four equating methods across three sampling conditions in samples from the "0.2 populations." Values for two linear and two curvilinear equating methods after matching on the selection variable are shown in the middle portion of Table 2. To the left are results after matching on the equating test; at the right, results in the representative samples are shown. Data for the two sampling conditions shown on the left and right come from the Livingston, Dorans and Wright (1990) study; data in the middle four columns were generated for the present study. It should be noted that only two equating methods are included for the matched-on-equating-test sampling condition: linear and equipercentile. This is because the Tucker and Levine linear methods and the chained and frequency estimation equipercentile methods converge in each case to a single equating result when perfect matching has been performed with the equating test.

Insert Table 2 about here

Figures 6a to 6d display data from Table 2. Comparison of the linear equatings under the two matching conditions (R and V versus M) show more accurate results in all four samples when matching has been performed with the selection variable. Results for equipercentile equatings nder the two matching conditions (D and Y compared with E) are mixed, with the selection variable clearly improving the results in only one sample, "ma." In the other three samples, the RMWSD values are similar or slightly larger for the selection-variable matched resı lts, whereas the bias values are smaller.

Insert Figure 6 about here

Comparisons within each equating method between the representative andmatched-on-selection-variable conditions show the verbal equatings to be about as accurate regardless of sampling condition, with all results in the acceptable range. The math equatings exhibit a somewhat different pattern. Tucker and frequency estimation results are improved under matching in the "ma" sample, but are about the same in the "mb" group. For the chained equipercentile method, representative-sample results were considerably more accurate in the "mb" sample, but similar in the "ma" sample.

Discussion

This research appears to confuse the issue about whether or not to match. The series of papers that motivated this study presented both positive and negative results with respect to matching on the equating test. The observational studies (Lawrence & Dorans, 1990; Schmitt, Cook, Dorans & Eignor, 1990) presented empirical support, while the simulation studies (Eignor, Stocking & Cook, 1990; Livingston, Dorans & Wright, 1990) presented results that questioned the efficacy of matching on the equating test. The present paper contains both positive and negative results with respect to matching. The negative results appeared in the Livingston, Dorans and Wright (1990) study: matching on the equating test hurts the performance of the item response theory, Levine and chained equipercentile equating methods, while it neither helps nor hurts the Tucker and frequency estimation equating methods. The reason that Tucker and frequency estimation are invariant with respect to matching on the equating test is that the mathematical assumptions underlying these methods are consistent with the logic of matching on the equating test. Since they are, in effect, statistical ways of matching, Tucker and frequency estimation obviate any need to match on equating test scores. So matching on equating test scores should be discontinued.

The positive results with respect to matching that were found in the present study were that matching on the selection variable, if known, is a correct thing to do. The results for both Tucker and frequency estimation were improved when equating was performed on samples matched with respect to the selection variable that had been used to construct the psuedopopulations in the Livingston, Dorans and Wright (1990) simulation study. In fact, by matching on the selection varaible, the effects of selection which had caused poor results for the Tucker and frequency estimation methods obtained using the equating test as either an anchor or matching variable were counteracted. In addition, matching on the selection variable did not have a detrimental effect on the performance of the Levine and chained equipercentile methods, producing results comparable to the results obtained under representative samples for these methods (see Table 2). Hence, the data suggest that matching on the selection variable, if known, as it was in this study, can improve equating results for some methods, e.g. those with a statistical foundation in selection theory: Tucker, frequency estimation, and, perhaps, the kernel equating method (Holland & Thayer, 1989). In addition, matching on the selection variable did not hurt the performance of any method in this study, the way that matching on the equating test did in Livingston, Dorans and Wright (1990).

Use of the selection variable as an anchor provided the most interesting results. Both the Tucker and frequency estimation procedures performed better, in most cases, with the selection variable as an anchor than they did when the common item equating test served as the anchor. It seems rather absurd that a verbal scaled score would be a better anchor for equating two math tests than would a mini-test in math. It is absurd until you think about it and realize that both the Tucker and frequency estimation methods, in essence, assume that the anchor they are using is, in effect, the selection variable, the variable along which the old and new form samples differ. In the math equatings, this selection variable was the verbal scaled score; for verbal, the selection variable was the math scaled score.

In contrast to the absurdly good performance of Tucker and frequency estimation, use of the selection variable as an anchor produced unreasonable and abominable results for Levine and chained equipercentile. (Comparably poor results would probably have occurred for IRT as well.) The poor results for Levine are easy to explain: The true score correlation between verbal and math is not unity, which id what the Levine model assumes about the true score correlation between the anchor test and the test to be equated. Chained equipercentile does poorly because the scaling relationship between verbal and math across psuedopopulations differs systematically. The simulated psuedopopulations were constructed in such a way that differences between populations at the mean along the selection variable were always larger than differences at the mean of the score to be equated. The simulation model in essence constructed data that the chained equipercentile model could not deal with if the selection variable was to be used as an anchor. Likewise, the simulation model had constructed data that prevented the Tucker and frequency estimation methods from obtaining reasonable results with the traditional equating test anchor.

The importance of constructing realistic simulations is one lesson to learn from this study. Livingston, Dorans and Wright (1990) went out of their way to avoid biasing results in favor of matching on the equating test by using a verbal score as a selection variable to set up psuedopopulations for math equatings, and vice versa for verbal equatings. Using the same data, we have shown matching to work well if matching occurs on the selection variable. Matching works, if you match on the right thing. We also show that the absurd -- using a verbal score as an anchor for a math equating -- works well with these data for certain models. That speaks to both the flexibility of the data, and the unrealistic nature of the simulation.

The most important lesson to be learned from this study is that equatings using anchor test designs in which the old and new form populations differ in ability can be improved if we can identify and properly use variables that describe the self-selection process underlying test-taking behavior. Perhaps equating research efforts should shift their focus towards attaining a better understanding of how naturally-occurring test populations occur. We have many models for relating test scores. Item response theory provides models for relating item scores to proficiency. We need better models of test-selection behavior in order to improve the quality of our equatings.

References

Angoff, W. H. (1982). Summary and derivation of equating methods used at ETS. In P. W. Holland & D. B. Rubin (Eds.), *Test equating* (pps. 55-69). New York: Academic Press.

Angoff, W. H. (1984). *Scales, norms and equivalent scores.* Princeton, NJ: Educational Testing Service.

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

Brennan, R. L. (1987). Introduction to problems, perspectives, and practical issues in equating. *Applied Psychological Measurement, 11,* 221-224.

Cook, L.L., Eignor, D. R., & Schmitt, A. P. (1989). *Equating Achievement tests using samples matched on ability.* Paper presented at the annual meeting of the Amercian Educational Research Association, San Francisco.

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.

Dorans, N. J. (1990a). (Ed.) Selecting samples for equating: To match or not to match. [Special issue]. *Applied Measurement in Education, 3.*

Dorans, N. J. (1990b). The equating methods and sampling designs. *Applied Measurement in Education. 3,* 3-17.

Eignor, D. R., Stocking, M. L., & Cook, L. L. (1990). Simulation results of the effects on linear and curvilinear observed- and true-score equating procedures of matching on a fallible criterion. *Applied Measurement in Education, 3,* 37-55.

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

Holland, P.W., & Thayer, D. T. (1989). *The kernel method of equating score distributions.* (RR-89-7). Educational Testing Service, Princeton, NJ.

Kolen, M. J. (1984). Effectiveness of analytic smoothing in equipercentile equating. *Journal of Educational Statistics, 9,* 25-44.

Kolen, M. J. (1990). Does matching in an equating work? A discussion. *Applied Measurement in Education, 3,* 97-104.

Lawrence, I. M., & Dorans, N. J. (1988). *A comparison of observed score and true score equating methods for representative samples and samples matched on an anchor test (RR-88-23).* Princeton, NJ: Educational Testing Service.

Lawrence, I. M., & Dorans, N. J. (1990). The effect on equating results of matching samples on an anchor test. *Applied Measurement in Education, 3,* 19-36.

Linn, R. L. (1989). (Ed.) *Educational Measurement* (3rd ed.). New York:Macmillan.

Livingston, S. A., Dorans, N. J., & Wright, N. K. (1990). What combination of equating and sampling works best? *Applied Measurement in Education, 3,* 73-95.

Petersen, N. S., Kolen, M. J., & Hoover, H. D. (1989). Scaling, norming, and equating. In R. L. Linn (Ed.) *Educational measurementt* (pps. 221-262). (3rd ed.). New York: Macmillan.

Petersen, N. S., Marco, G. L., & Stewart, E. E. (1982). A test of the adequacy of linear score equating models. In P. W. Holland & D. R. Rubin (Eds.), *Test equating* (pps. 71-135). New York: Academic Press.

Potthoff, R. F. (1982). Some issues in test equating. In P. W. Holland & D. B. Rubin (Eds.), *Test equating*. (pps.201-242). New York, NY: Academic Press.

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

Schmitt, A. P., Cook, L. L., Dorans, N. J., & Eignor, D. R. (1990). The sensitivity of equating results to different sampling strategies. *Applied Measurement in Education, 3,* 53-71.

Skaggs, G. (1990). To match or not to match samples on ability for equating: A discussion of five articles. *Applied Measurement in Education, 3,* 105-113.

Skaggs, G., & Lissitz, R. W. (1986). IRT test equating: Relevant issues and a review ofrecent research. *Review of Educational Research, 56,* 495-529.

Table 1. Bias and Root Mean-Weighted Square Difference (RMWSD) for Equated Scores (on the College Board 200-to-800 Scale) Based on Different Equating Methods in Representative Samples

		Anchor					
		External Equating Test[a]				Selection Variable[b]	
Old Form Sample		Tucker	Frequency Estimation	Chained Equipercentile	Levine	Tucker	Frequency Estimation
va 0.4	Bias	-9.6	-9.0	-4.1	-1.7	-2.4	-2.9
	RMWSD	10.5	9.9	6.3	2.8	3.3	3.7
va 0.2	Bias	-3.8	-3.4	-1.3	-0.2	-0.6	-1.2
	RMWSD	4.7	4.4	4.4	2.3	2.3	2.6
va 0.0	Bias	-1.0	-0.9	-1.1	-1.0	-1.6	-1.6
	RMWSD	2.4	1.8	2.7	2.4	2.8	1.9
vb 0.4	Bias	-5.8	-5.0	+0.9	+3.6	-0.6	-3.5
	RMWSD	6.8	5.6	3.4	5.3	2.9	5.5
vb 0.2	Bias	+0.3	+0.5	+2.8	+3.7	+2.6	+1.7
	RMWSD	3.5	3.1	4.4	4.3	3.8	3.5
vb 0.0	Bias	+2.0	+1.8	+1.9	+1.9	-0.1	-0.2
	RMWSD	3.0	2.4	2.8	2.9	2.2	1.2
ma 0.4	Bias	-11.9	-11.1	-4.3	-0.5	-1.9	-1.5
	RMWSD	12.2	11.7	5.1	3.6	4.8	5.0
ma 0.2	Bias	-6.9	-6.5	-3.7	-2.2	-1.3	-1.8
	RMWSD	7.1	6.9	4.4	3.8	3.7	4.0
ma 0.0	Bias	-0.8	-0.8	-0.2	-0.2	-2.5	-3.0
	RMWSD	1.9	1.8	1.9	1.8	3.0	3.6
mb 0.4	Bias	-9.0	-8.7	-1.3	+4.4	-6.0	-7.0
	RMWSD	9.5	9.7	6.0	10.4	7.5	8.3
mb 0.2	Bias	-4.5	-4.4	-0.6	+1.7	-5.2	-6.1
	RMWSD	4.9	5.5	2.7	4.9	6.4	8.0
mb 0.0	Bias	-1.3	-1.3	-1.0	-1.1	-3.1	-3.2
	RMWSD	2.1	1.8	1.5	1.9	3.9	4.3

Note. The equipercentile equating in the full population was used as the criterion equating.

[a] Data taken from Livingston, Dorans and Wright (1990)

[b] Results for Chained Equipercentile and Levine are excluded because they were very large with RMWSD valued exceeding 50 in some cases.

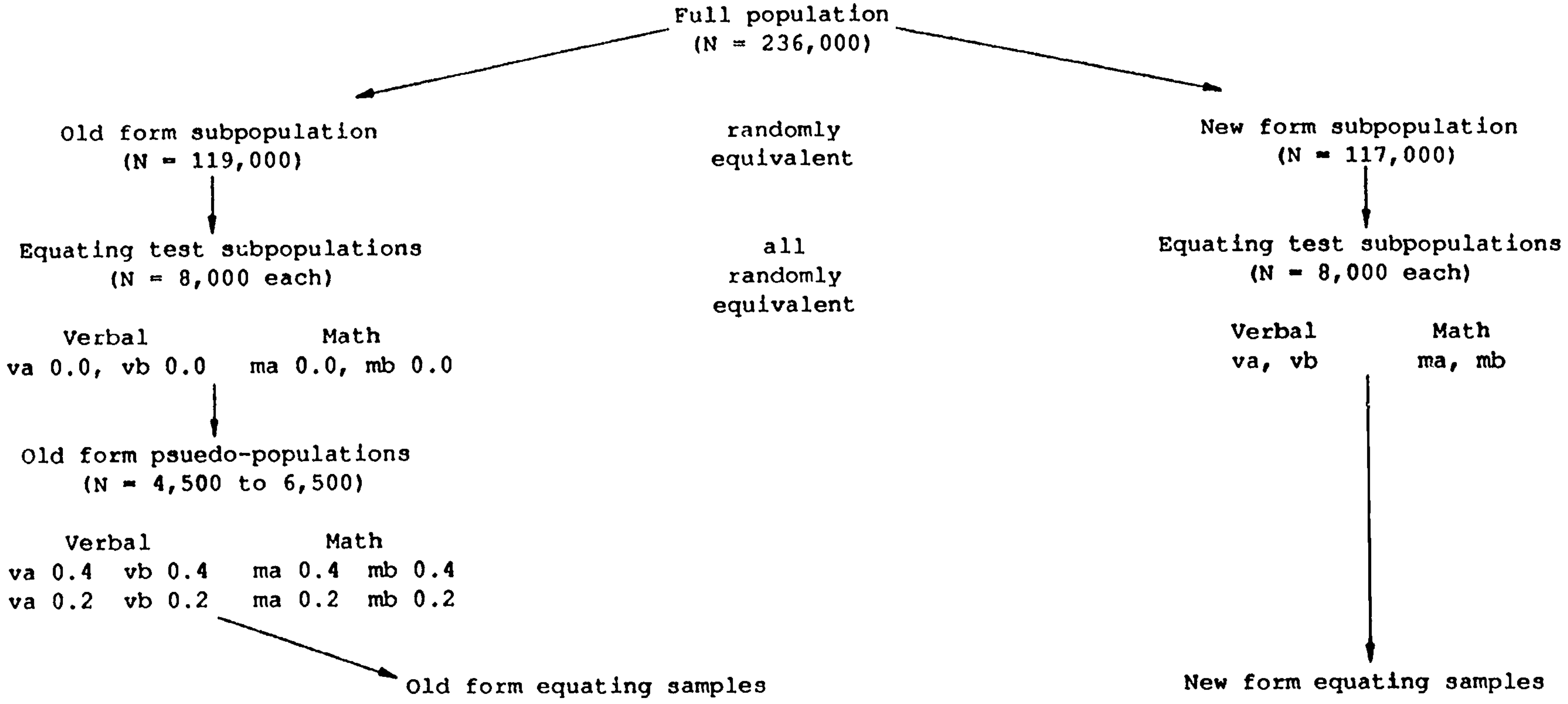
Table 2. Bias and Root Mean-Weighted Square Difference (RMWSD) for Equated Scores (on the College Board 200-to-800 Scale) Based on Different Equating Methods Using Matched Samples or Representative Samples

		Sampling Condition									
		Matched on Equating Test[a]		Matched on Selection Variable				Representative[a]			
		Linear	Equipercentile	Linear		Equipercentile		Linear		Equipercentile	
Old Form Sample				Tucker	Levine	Chained	Frequency Estimation	Tucker	Levine	Chained	Frequency Estimation
va 0.2	Bias	-3.6	-3.5	-1.6	-1.3	-1.3	-1.3	-3.8	-0.2	-1.3	-3.4
	RMWSD	4.4	4.8	2.8	2.6	4.7	3.8	4.7	2.3	4.4	4.4
vb 0.2	Bias	-2.7	-2.7	+1.4	+1.8	+1.7	+1.2	+0.3	+3.7	+2.8	+0.5
	RMWSD	3.5	3.3	2.7	2.9	4.3	3.4	3.5	4.3	4.4	3.1
ma 0.2	Bias	-5.8	-5.7	-3.2	-3.1	-2.8	-3.2	-6.9	-2.2	-3.7	-6.5
	RMWSD	6.2	6.4	3.6	3.5	4.1	4.4	7.1	3.8	4.4	6.9
mb 0.2	Bias	-5.1	-5.0	-3.0	-2.5	-2.5	-2.8	-4.5	+1.7	-0.6	-4.4
	RMWSD	5.7	5.9	5.2	4.9	6.3	6.6	4.9	4.9	2.7	5.5

Note The common-item equating test was used as the anchor in all equatings. The equipercentile equating in the full population was used as the criterion equating.

[a] Data taken from Livingston, Dorans and Wright (1990).

FIGURE 1. Design of study[a]



Full population
(N = 236,000)

Old form subpopulation
(N = 119,000)

Equating test subpopulations
(N = 8,000 each)

Verbal: va 0.0, vb 0.0
Math: ma 0.0, mb 0.0

Old form psuedo-populations
(N = 4,500 to 6,500)

Verbal: va 0.4 vb 0.4, va 0.2 vb 0.2
Math: ma 0.4 mb 0.4, ma 0.2 mb 0.2

Old form equating samples

randomly
equivalent

all
randomly
equivalent

New form subpopulation
(N = 117,000)

Equating test subpopulations
(N = 8,000 each)

Verbal: va, vb
Math: ma, mb

New form equating samples

	Representative (N = 3,000 each)	Matched on equating test (N = 3,000 each)	**Matched on selection variable (N = 2,000 each)**	Representative (N = 3,000 each)
Verbal:	va 0.4	(va 0.4)[b]	---[c]	va
	va 0.2	va 0.2	**va 0.2**	
Verbal:	vb 0.4	(vb 0.4)[b]	---[c]	vb
	vb 0.2	vb 0.2	**vb 0.2**	
Math:	ma 0.4	(ma 0.4)[b]	---[c]	ma
	ma 0.2	ma 0.2	**ma 0.2**	
Math:	mb 0.4	(mb 0.4)[b]	---[c]	mb
	mb 0.2	mb 0.2	**mb 0.2**	

[a] Samples in bold face selected for this study; all other populations and samples selected for Livingston, Dorans and Wright (1990).

[b] Not included in this study

[c] Not available for this study

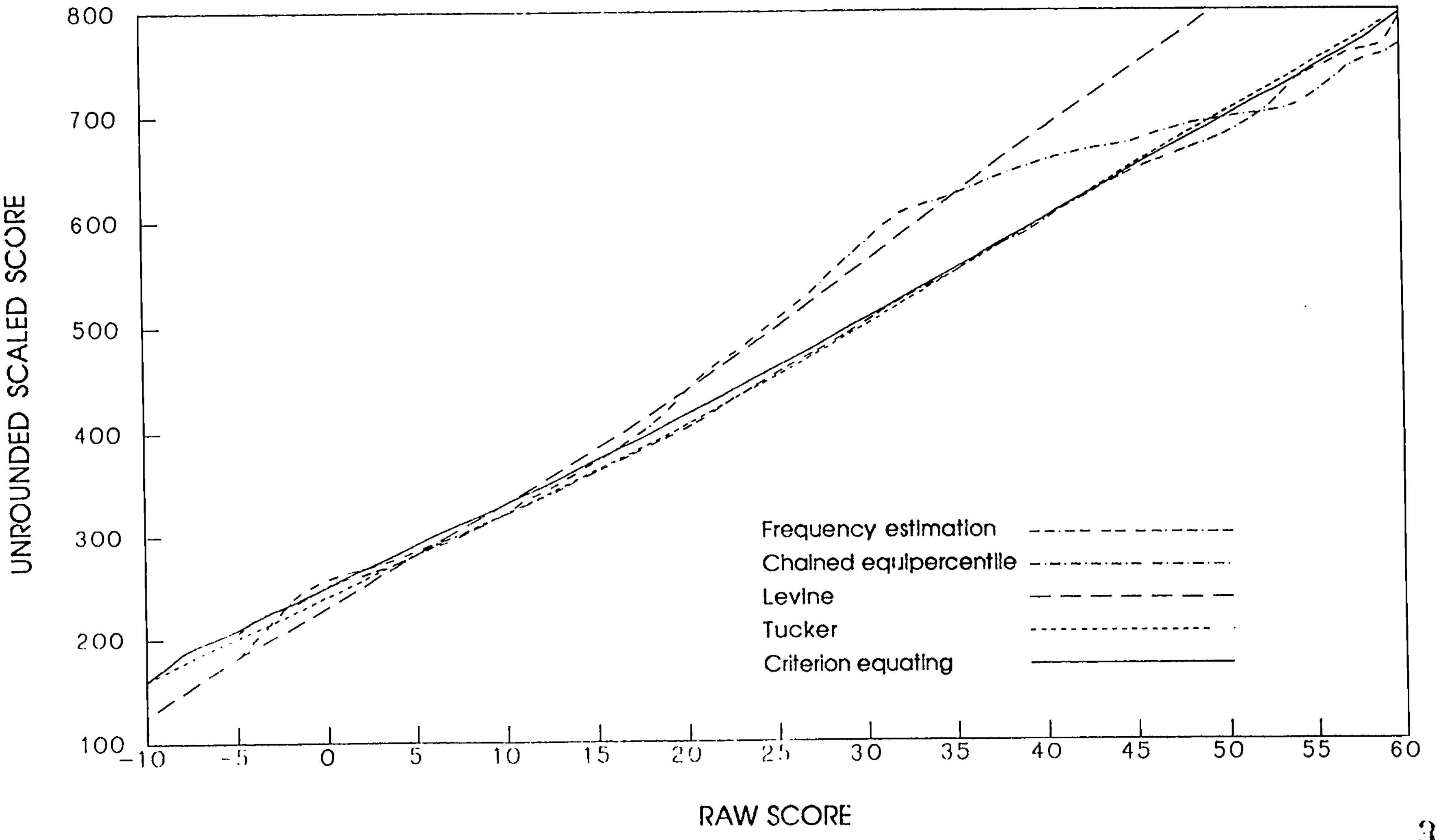


FIGURE 2. The criterion equating function and equating functions for math scores from the "mb 0.4 population" based on four equating methods in which the selection variable served as anchor.

VERBAL

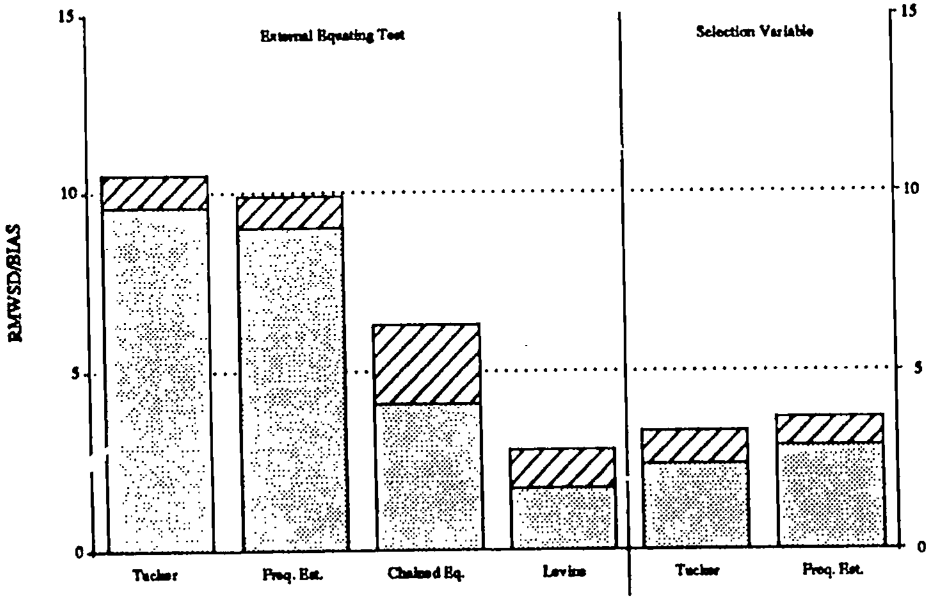


Figure 3a. Bias and RMWSD in equating the verbal scores through the equating test versus through the selection variable, sampling from the °0.4 vs population.

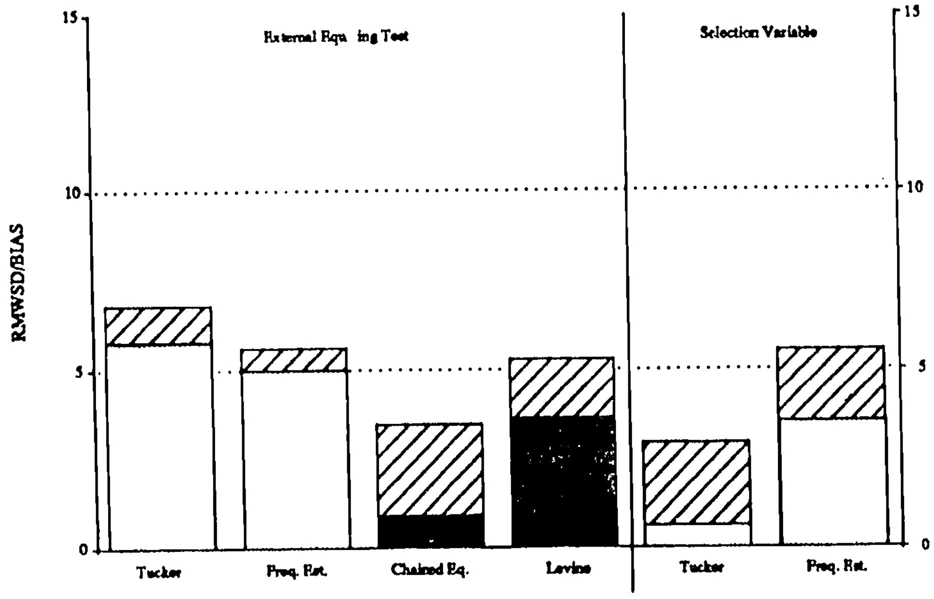


Figure 3b Bias and RMWSD in equating the verbal scores through the equating test versus through the selection variable, sampling from the °0.4 vb population.

MATH

☒ RMWSD
■ positive bias
□ negative bias

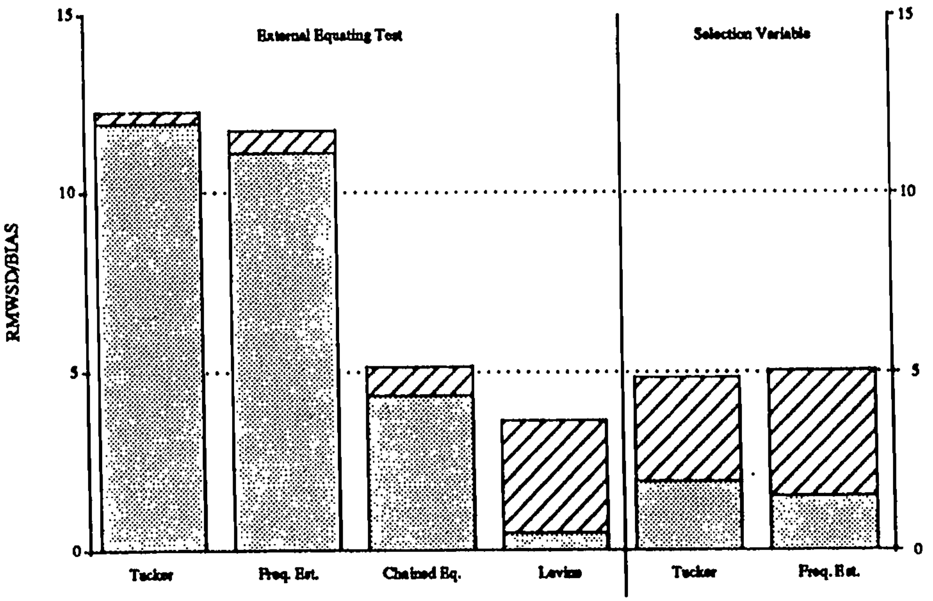


Figure 3c. Bias and R..IWSD in equating the verbal scores through the equating test versus through the selection variable, sampling from the °0.4 ms population.

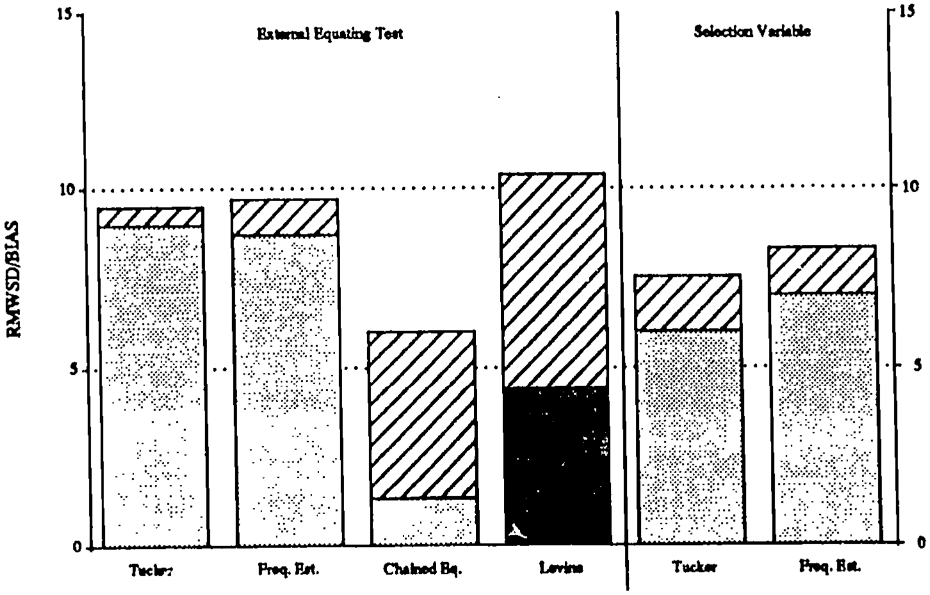


Figure 3d. Bias and RMWSD in equating the verbal scores through the equating test versus through the selection variable, sampling from the °0.4 mb population.

VERBAL

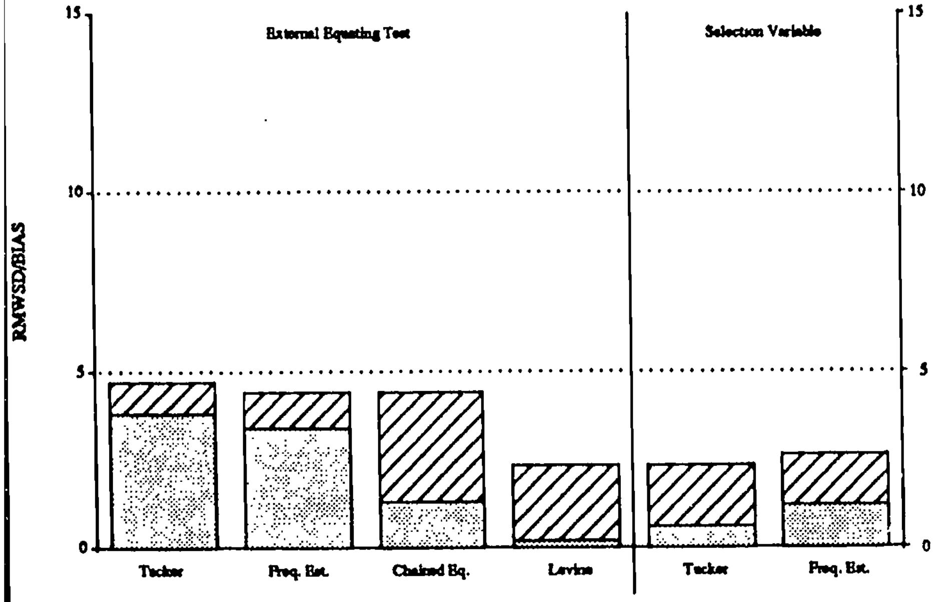


Figure 4a. Bias and RMWSD in equating the verbal scores through the equating test versus through the selection variable, sampling from the °0.2 va population.

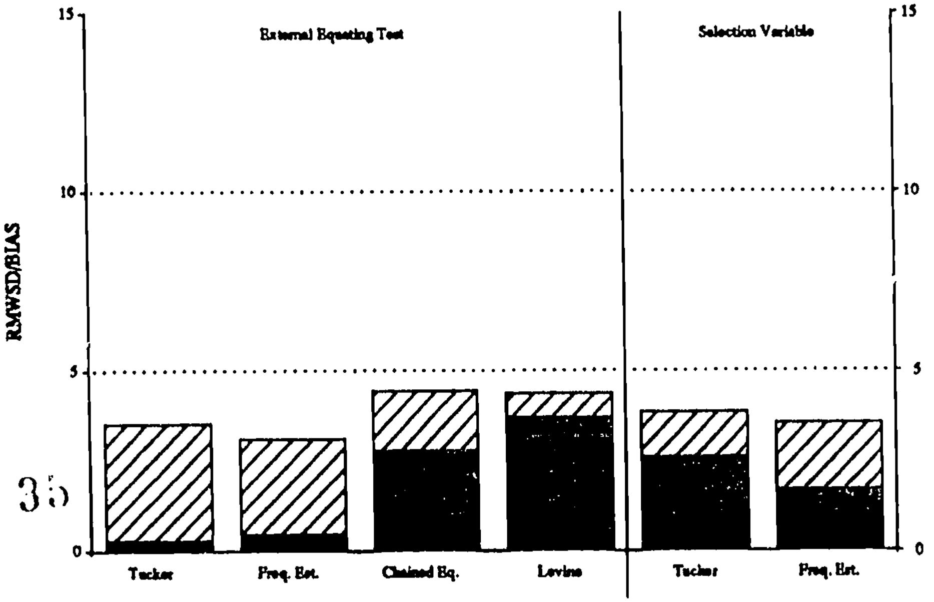


Figure 4b. Bias and RMWSD in equating the verbal scores through the equating test versus through the selection variable, sampling from the °0 2 vb populatjon.

MATH

☐ RMWSD
▨ positive bias
☐ negative bias

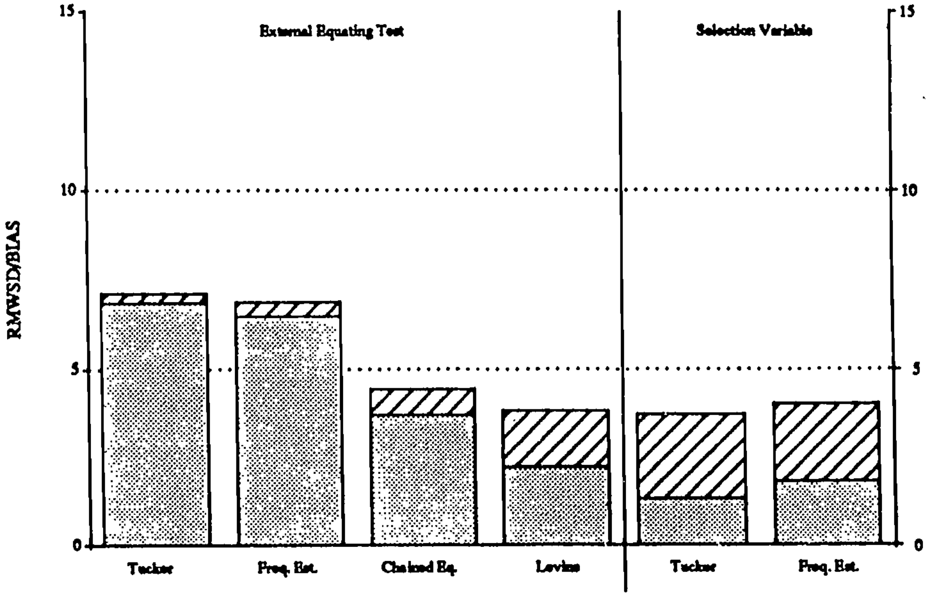


Figure 4c. Bias and RMWSD in equating the verbal scores through the equating test versus through the selection variable, sampl ing from the °0.2 ma population.

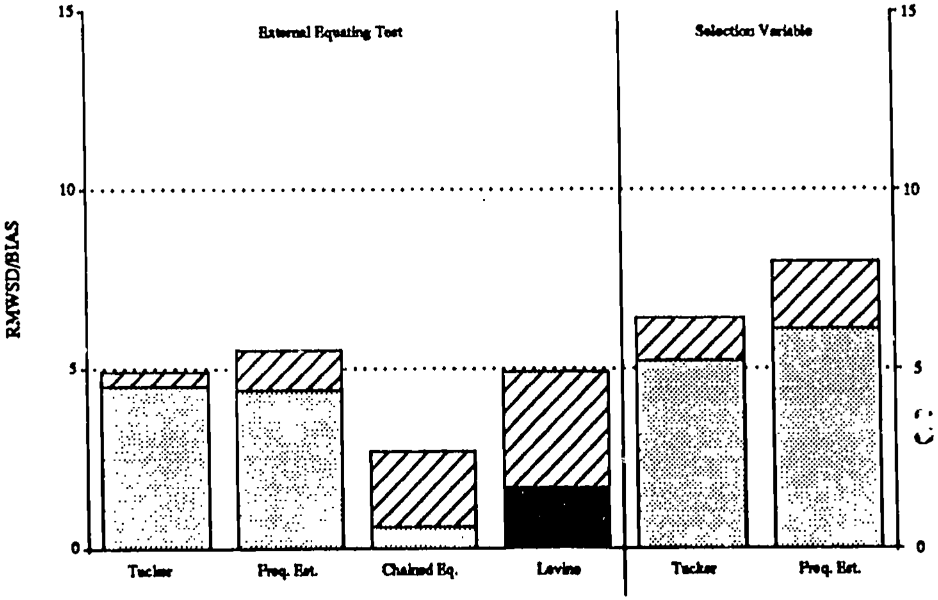


Figure 4d. Bias and RMWSD in equating the verbal scores through the equating test versus through the selection variable, sampl-.q from the °0.2 mb population.

VERBAL

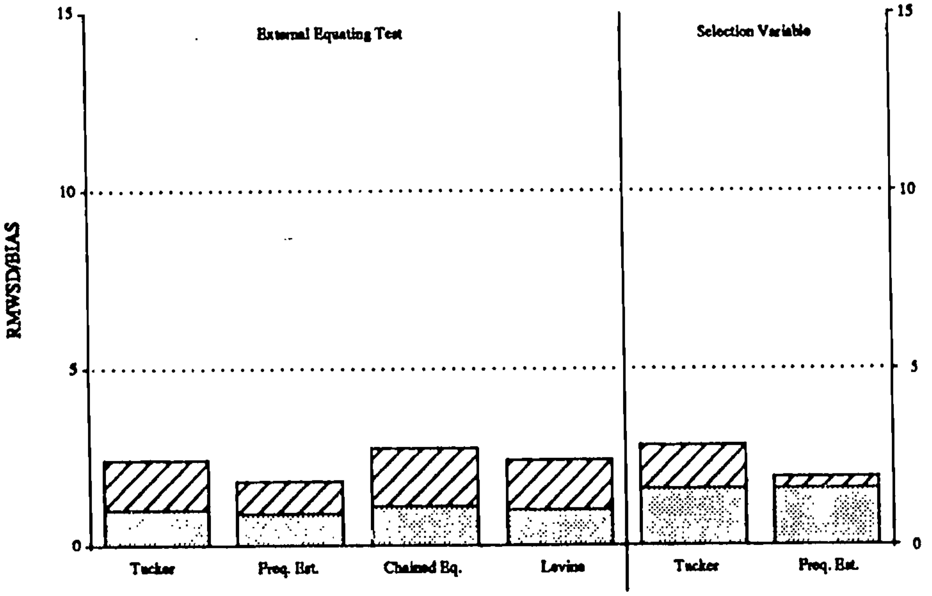


Figure 5b. Bias and RMWSD in equating the verbal scores through the equating test versus through the selection variable, sampling from the °0.0 va population.

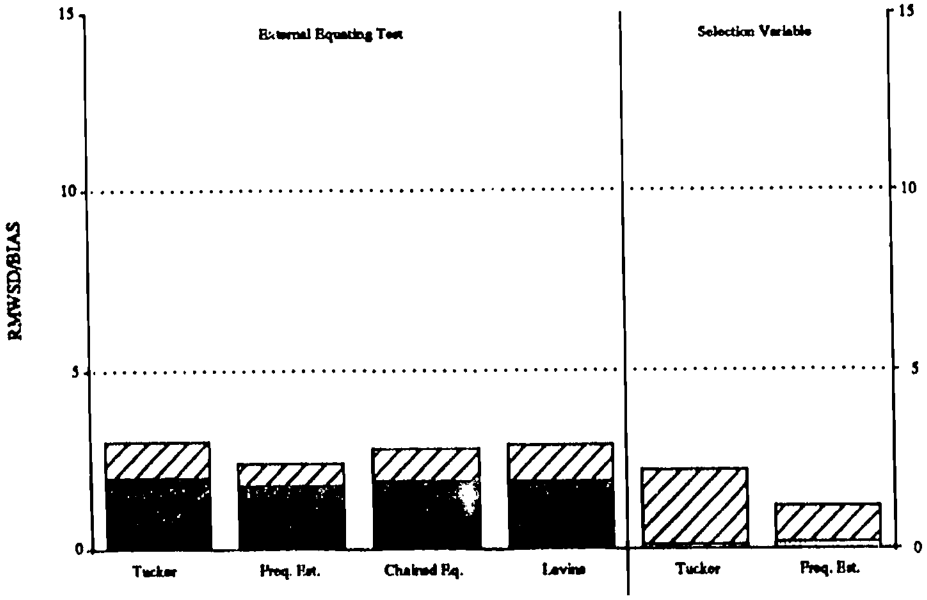


Figure 5a. Bias and RMWSD in equating the verbal scores through the equating test w rsus through the selection variable, sampling from the °0 0 vb population.

MATH

□ RMWSD
■ positive bias
▨ negative bias

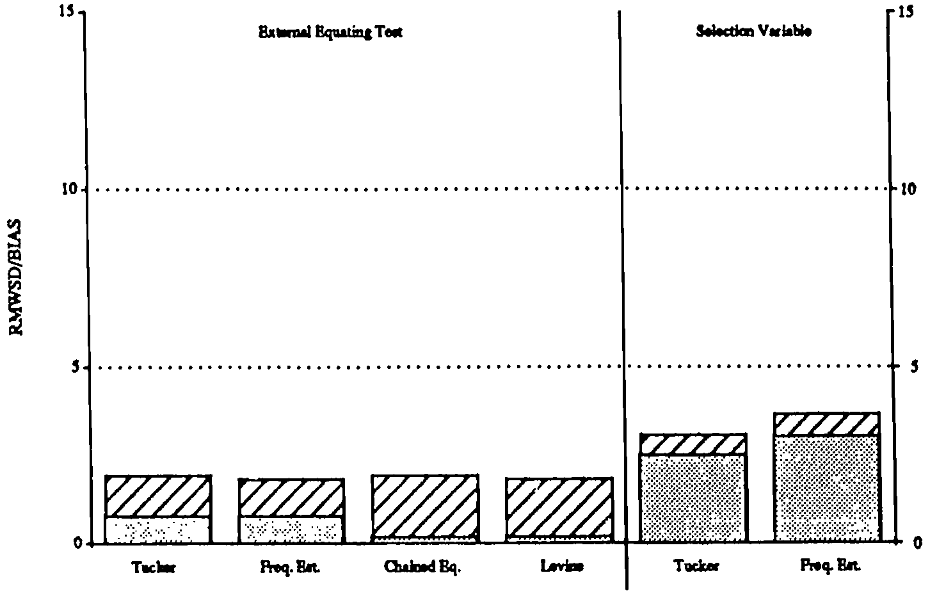


Figure 5c. Bias and RMWSD in equating the verbal scores through the equating test versus through the selection variable, sampling from the °0.0 ma population.

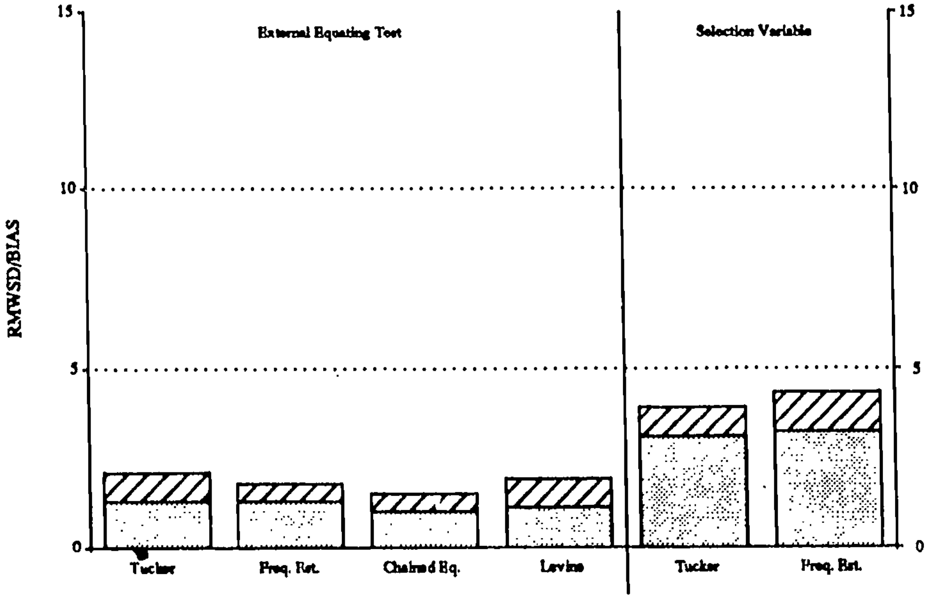


Figure 5d. Bias and RMWSD in e ,uating the verbal scores through the equating test versus through the selection variable, sampling from the °0.0 mb population.

VERBAL

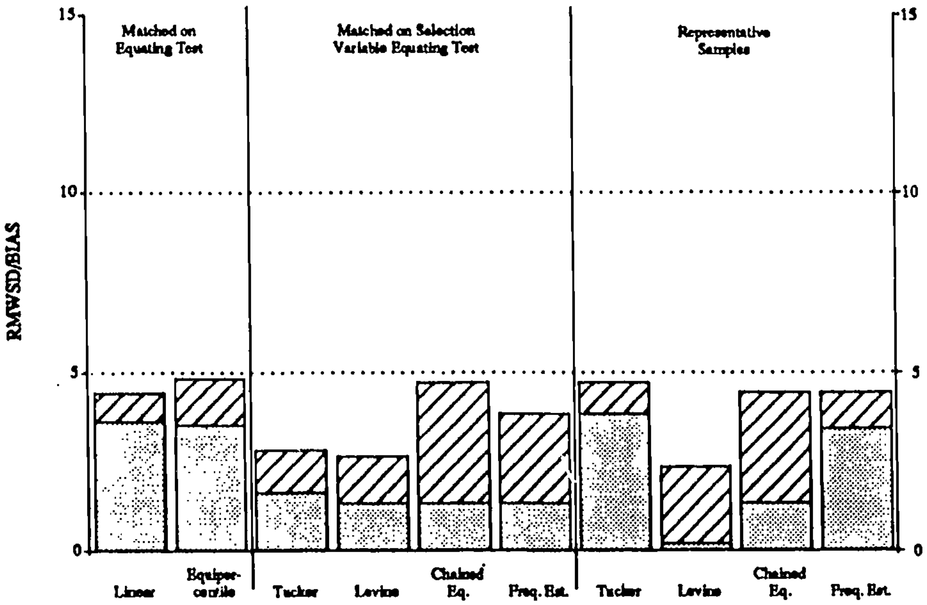


Figure 6a. Bias and RMWSD in equating the verbal scores through the equating test versus through the selection variable, sampling from the '0.2 va population.

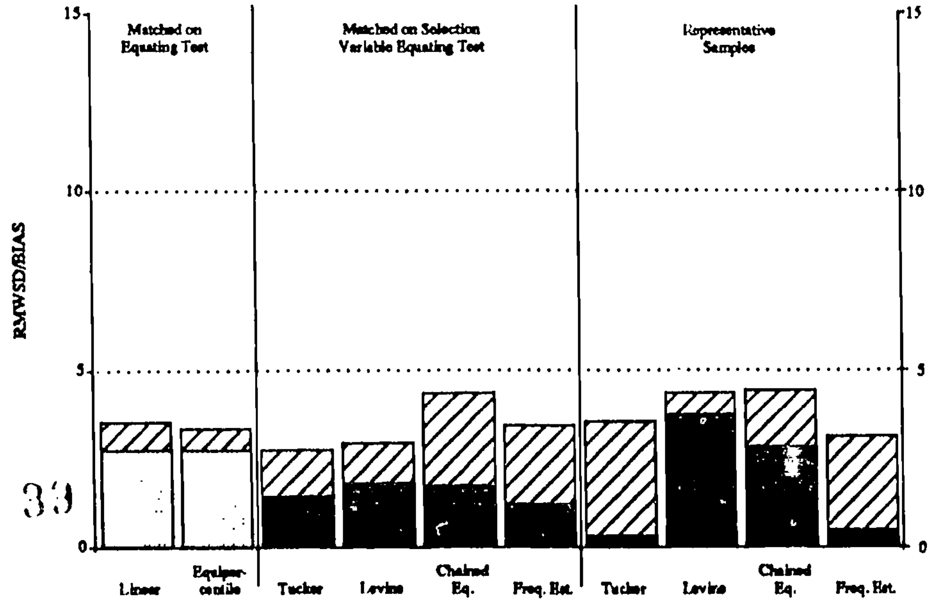


Figure 6b. Bias and RMWSD in equating the verbal scores through the equating test versus through the selection variable, sampling from the '0.2 vb population.

MATH

☐ RMWSD
■ positive bias
☐ negative bias

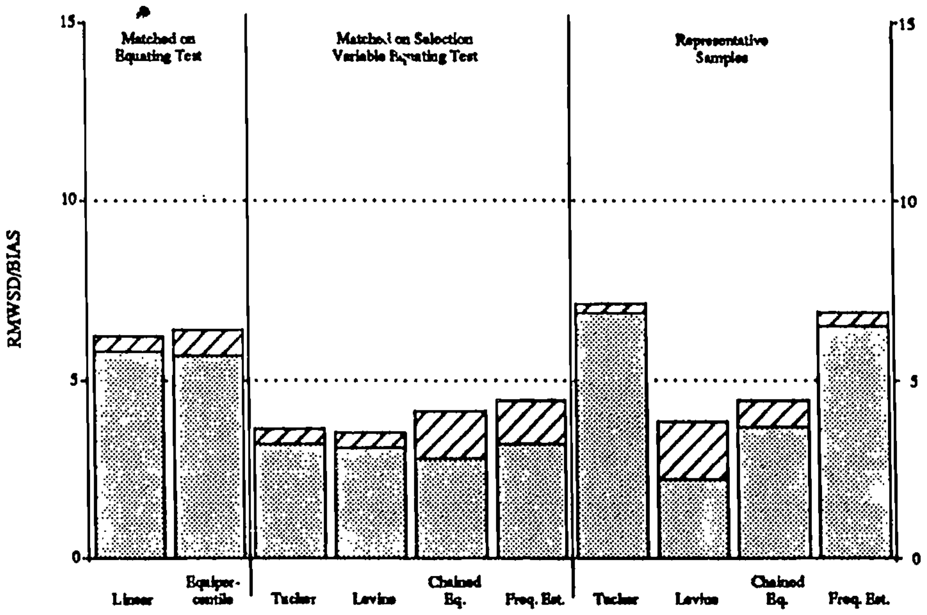


Figure 6c. Bias and RMWSD in equating the verbal scores through the equating test versus through the selection variable, sampling from the '0.2 ma population.

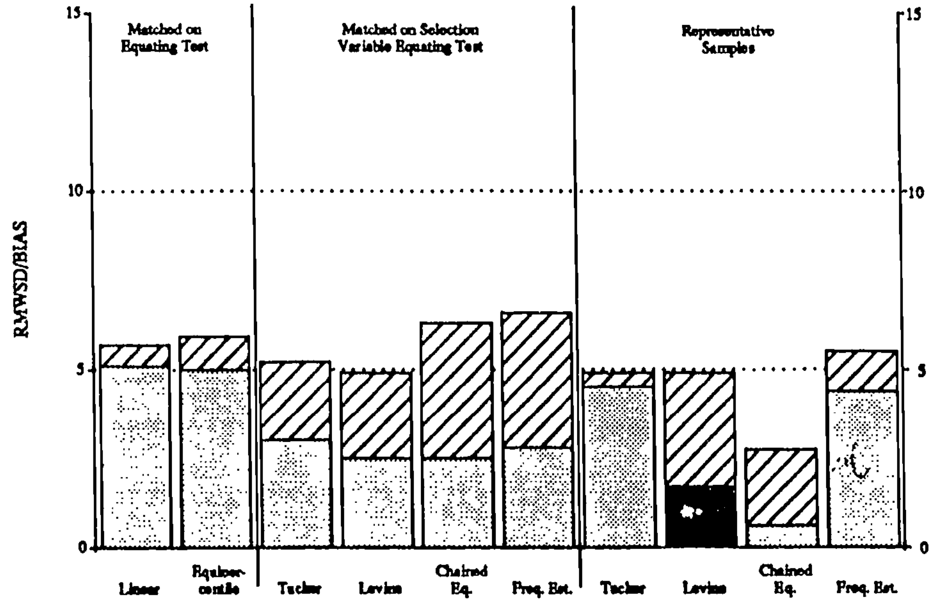


Figure 6d. Bias and RMWSD in equating the verbal scores through the equating test versus through the selection variable, sampling from the '0.2 mb population.

BEST COPY AVAILABLE