DOCUMENT RESUME

ED 384 658 TM 023 951

AUTHOR	Angoff, William H.
TITLE	The Determination of Empirical Standard Errors of Equating the Scores on SAT-Verbal and SAT-Mathematical.
INSTITUTION	Educational Testing Service, Princeton, N.J.
REPORT NO	ETS-RR-91-54
PUB DATE	Oct 91
NOTE	12p.
PUB TYPE	Reports - Evaluative/Feasibility (142)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	*College Entrance Examinations; *Equated Scores; *Error of Measurement; Higher Education; High School Students; Mathematics Tests; Scoring; *Statistical Analysis; Test Format; *Testing Programs; Verbal Tests
IDENTIFIERS	Angoff Methods; Empirical Research; *Scholastic Aptitude Test; *Variance (Statistical)

ABSTRACT

An attempt was made to evaluate the standard error of equating (at the mean of the scores) in an ongoing testing program. The interest in estimating the empirical standard error of equating is occasioned by some discomfort with the error normally reported for test scores. Data used for this evaluation came from the Admissions Testing Program of the College Board. The method used depends on the fact that about half the examinees take the Scholastic Aptitude Test (SAT) twice or more. The calculation of the standard error of equating SAT verbal and mathematical scores in this study makes use of the variance of the mean gains over the course of 12 to 17 years for which comparable data are available, separately by pattern of repetition. It is reasoned that because each of the means used to calculate a mean gain is based on data for a different form of the test, the variance of a number of these means would be attributable to some extent to the variation associated with equating error, and that the variance of the errors of the mean gains would equal the variance of the errors on the first occasion of testing plus the variance of the errors on the second occasion, or twice the variance of errors on either occasion. In the example, standard errors of equating were quite small. One table presents the standard errors. (SLD)

RR-91-54

RESEARCH REPORT

THE DETERMINATION OF EMPIRICAL STANDARD ERRORS OF EQUATING THE SCORES ON SAT-VERBAL AND SAT-MATHEMATICAL

William H. Angoff

Educational Testing Service
Princeton, New Jersey
October 1991

The Determination of Empirical Standard Errors of Equating the Scores on SAT-Verbal and SAT-Mathematical

William H. Angoff

The author wishes to express his appreciation to Daniel Eignor, Eugene Johnson, Charles Lewis, and Nancy Wright for their very helpful reviews of this paper.

This research was supported in part by the ETS Office of Corporate Quality Assurance.

The present study is an attempt to evaluate the standard error of equating (at the mean of the scores) in an ongoing testing program. Unlike the conditions typically specified for the calculation of equating error (see Angoff, 1984; Lord, 1950), those that prevail in an actual testing program vary considerably and usually fail to meet the specified conditions. For example, the methods of equating that are applied in practice--in the equating of forms of the SAT, for example, in which various forms of linear and curvilinear methods are used--are not always precisely the same from one administration to another. Further, the numbers of cases used for equating frequently vary from one time to another, and the internal statistics -- e.g., the average and the dispersion of ability of the groups used for equating and the correlation of the equating test and the tests to be equated -- also vary somewhat from one administration to another. And perhaps most important, the random conditions specified in the formulas for standard errors typically do not hold for an equating done in a testing program. As a consequence, the size of the standard error is ordinarily impossible to determine formally, and the best one can do is to make an educated guess at its value. Indeed, the very term, "standard error", under conditions prevailing in a testing program is not entirely appropriate inasmuch as the concept assumes certain specified constraints; and if these conditions do not exist, then the concept loses much of its meaning. In spite of this, there is some validity in the search for a measure of the amount of variation in the reported score caused by the effort to equate the forms of the tests.

The interest in estimating the empirical standard error of equating is occasioned by some discomfort with the error normally reported for test scores. Typically, the error we associate with the test score, i.e., the error reported in the interpretive literature for the test, is the raw score

standard error of measurement converted into the scaled score terms used in reporting scores to the users. But, however converted, this number is still only the standard error of measurement, and does not contain the component of error due to equating. It is this latter component that we now seek to evaluate.

The data used for this evaluation come from the Admissions Testing Program of the College Board. The method used depends on the fact that about half of the candidates who take the SAT take the test twice or more, once (typically) in March, May, or June of their junior year in high school, and again (also typically) in October, November, December, or January of their senior year. However, these patterns of repetition and the numbers of students choosing to take the tests in these patterns of repetition have varied over the years. Over all, the mean gains have averaged between 15 and 25 scaled score points, with greater gains for students who take the tests several times than for those who take them fewer times and with declining gains on successive testings for multiple repeaters. The standard deviations of gains run about 45 to 50 for SAT-Verbal and about 50 to 55 for SAT-Mathematical.

The calculation of the standard error of equating SAT-Verbal and SAT-Mathematical scores (at the mean of the scores) in this study makes use of the variance of the mean gains over the course of 12-17 years for which comparable data are available, separately by pattern of repetition. It is reasoned that because each of the means used to calculate a mean gain is based on data for a different form of the test--and recognizing that the forms differ over the years

--the variance of a number of these means would be to some extent attributable to the variation associated with equating error; also that the variance of the errors of the mean gains would equal the variance of the errors on the first occasion of testing plus the variance of the errors on the second occasion of testing, or twice the variance of errors on either occasion (inasmuch as we have no reason to believe that the variance of errors on the second occasion are any different from those on the first). It is also reasoned that the errors associated with these mean gains are independent of the errors of measurement and that the errors of measurement in the means are essentially zero; the mean gains are observed in samples numbering, typically, in the tens of thousands -- in these data, although they vary considerably, the sample sizes average about 32,500 cases -- and errors of measurement are more than likely to vanish with samples of this size.

It is likely that there are additional sources of variance that account for the variance of the mean gains. For example, there is the error of sampling. But again, with samples of the size dealt with here errors attributable to sampling variation would add very little to the estimate of equating error. It is also recalled that the equating error associated with a score varies (increases) as a function of the distance of the score from the mean. In the case of the repeaters simple generalizations regarding their mean levels of performance are difficult to make: First, there are great variations in the size of the mean scores of repeaters. Second, those who repeat the test fewer times appear to score higher than those who repeat the test several times. (This is to be expected inasmuch as it is reasonable to speculate that many, perhaps most, of those who repeat the test do so because

they are not pleased with their previous performance and hope to do better on another try, and those who engage in fewer repetitions are more satisfied with their previous performance than those who continue to repeat the test. It should be noted, incidentally, that repeaters are not necessarily lower-scoring candidates. For example, many of the once-only repeaters repeat mainly because their colleges of application, typically the more selective colleges, expect them to take the test a second time. Not only these once-only repeaters, but also other groups of repeaters, score higher, on the average, than the one-time takers.) There may be other sources of variance that, ideally, should be controlled for, or eliminated, but they are likely to be difficult if not impossible to evaluate and remove from the measure of equating

error. On the other hand, it is expected that these additional sources of variance, which are not eliminated here, should be very small and should add minimally to the size of the "pure" error of equating. The error reported here may therefore be considered a conservative, or upper-bound estimate.

It was in fact possible to eliminate one source of variation. It was thought that the mean gains might vary systematically as a function of the particular repeater pattern, inasmuch as the patterns differ with respect to the amount of time elapsing between the first and second testing. Accordingly, it was thought advisable to evaluate the error separately by pattern of repetition. This decision was later supported by the data; the mean gains did vary with the amount of time between the two testings.

As indicated earlier, the calculation of the standard error of equating (at the mean) tabled here makes use of the variance of mean scaled score

gains. It is reasoned that the errors in both the first and in the second mean scores are largely errors of equating. Therefore the variance of the gains should be equal to twice the error of equating, and the square root of one-half of this variance can be taken as the empirical standard error of equating, which is estimated and shown in Table 1 separately by pattern of repetition.

Table 1

Empirical Standard Errors of Equating of SAT Scaled Scores
Separately by Pattern of Test Repetition

Pattern of Repetition	No. of Months Elapsing	Number of Mean Gains Available Per Pattern	Average of Mean Gains		Variance of Mean Gains		Standard Error of Equating	
			Verbal	Math	Verbal	Math	Verbal	Math
March-October	7	12	16.33	14.84	21.55	18.72	3.28	3.06
March-November	8	16	15.40	15.89	24.85	31.00	3.52	3.94
March-December	9	16	18.56	21.24	29.64	14.47	3.85	2.69
March-January	10	16	21.12	22.58	22.48	48.35	3.35	4.92
May-October	5	13	13.70	14.67	21.77	30.02	3.30	3.87
May-November	6	17	13.94	14.42	16.66	52.74	2.89	5.14
May-December	7	16	16.51	20.42	23.73	35.42	3.44	4.21
May-January	8	17	20.95	23.02	23.05	60.43	3.39	5.50
June-October	4	16	10.92	10.26	20.79	21.36	3.32	3.27
June-November	5	16	12.65	11.79	22.67	27.94	3.37	3.74
June-December	6	15	14.01	18.63	18.29	18.12	3.02	3.01
June-January	7	15	18.89	21.93	22.71	37.24	3.37	4.32
Mean over Patterns			16.08	17.47	22.35	32.98	3.33	3.97
Standard Deviation			3.12	16.91	3.09	13.98	.229	.728

The standard errors of equating reported in Table 1 appear to average about 3-1/3 scaled score points for Verbal and about 4 points for Math. However, the variability of the Math estimates across the 12 determinations

(repeater patterns) is much larger than that of the Verbal estimates; the standard deviation for Math is about 3.2 times greater than the standard deviation for Verbal. This is, as expected, consistent with the relative variabilities of the mean gains; the standard deviation of the average mean Math gain is 5.4 times as large as the standard deviation of the average mean Verbal gain.

It is observed that the variability of the mean gains in Math is much greater when January is the second administration than when October, November, or December is the second administration. It is possible that this is so because the January administration has served a variety of functions for the candidate populations over the course of these years and was therefore particularly sensitive to self-selection effects over these years. If we omit the data for the January repeaters, the average standard error of equating of SAT-Math drops from 3.97 to 3.66 and the variability of these estimates drops from .728 to .707, still much greater than Verbal, but now by a factor of 3.1 to 1 (instead of 3.2 to 1).

It is interesting that in the case of both Verbal and Math the average mean gain is closely associated with the amount of elapsed time between the first of these measurements and the second; it was for this reason that the standard errors of equating were calculated separately by pattern of repetition.. (It should be noted in this connection that although these measurements are described here as "first" and "second", they are so only in the sense that the "second" measurement followed the "first". In fact, there may have been several other measurements, some before the "first", some after

the "second", and perhaps some intervening between the "first" and the "second".)

In general these standard errors of equating are quite small -- 3.3 scaled score points for Verbal and 4.0 for Math -- representing about 3%-3 1/2% of the standard deviations of scaled scores and probably smaller, considering the extraneous sources of variance mentioned above. This being the case, they should not add appreciably to the standard errors of measurement already reported for SAT scores. If, for example, the (average) standard error of measurement for the Verbal score is taken as 31.5 scaled score points in scaled score terms, the standard error of the sum of measurement error and equating error, assuming an average standard error of equating of 3.3 points as calculated and reported here, would be 31.7 scaled score points (the square root of the sum of the variance of the errors of measurement and the variance of the errors of equating), hardly a significant inflation of the standard error of measurement taken by itself.

A final note of caution in interpreting these data is appropriate: The standard errors calculated here apply to the means of the distributions of first and second scores considered in the mean gains. As already noted, as one moves from the means of the distributions to the extremes, the equating errors increase in size. The standard errors of measurement, however, decrease in size as one departs from the mean. What the size of the error of measurement-plus-equating may be at any particular point other than those considered here would have to be evaluated in a study beyond the scope of the present effort.

References

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

Lord, F. M. (1950). Notes on comparable scales for test scores. Research Bulletin, No. 48. Princeton, NJ: Educational Testing Service.