Research Report No. 2002-11

The College Board

The Recentering of SAT® Scales and Its Effects on Score Distributions and Score Interpretations

Neil J. Dorans

College Board Research Report No. 2002-11
ETS RR-02-04

The Recentering of SAT® Scales and Its Effects on Score Distributions and Score Interpretations

Neil J. Dorans

College Entrance Examination Board, New York, 2002

Neil J. Dorans is principal measurement statistician at Educational Testing Service.

The College Board: Expanding College Opportunity

The College Board is a national nonprofit membership association dedicated to preparing, inspiring, and connecting students to college and opportunity. Founded in 1900, the association is composed of more than 4,200 schools, colleges, universities, and other educational organizations. Each year, the College Board serves over three million students and their parents, 22,000 high schools, and 3,500 colleges, through major programs and services in college admissions, guidance, assessment, financial aid, enrollment, and teaching and learning. Among its best-known programs are the SAT®, the PSAT/NMSQT®, and the Advanced Placement Program® (AP®). The College Board is committed to the principles of excellence and equity, and that commitment is embodied in all of its programs, services, activities, and concerns.

For further information, contact www.collegeboard.com.

Additional copies of this report (item #993941) may be obtained from College Board Publications, Box 886, New York, NY 10101-0886, 800 323-7155. The price is $15. Please include $4 for postage and handling.

Printed in the United States of America.

Acknowledgments

Many people contributed to the recentering work described here. Special recognition goes to Nancy Feryok who contributed throughout the data preparation and data analysis phases of the recentering analysis. Nancy Wright played a major role in organizing and coordinating the data preparation phases of this study. Paul Holland and Dorothy Thayer played critical roles in the smoothing and continuization phases of this study. Paul Holland, Charles Lewis, and Gary Marco provided technical advice at various points. Linda Cook and Larry Hecht played important roles in convincing the College Board of the merit of recentering. Bernadette McIntosh provided graphics support. Miriam Feigenbaum coordinated preparation of the data displays, which were developed by Ted Blew and myself, based on suggestions provided by Samuel Livingston and Larry Hecht. Reviews of earlier versions of this paper by Nancy Burton, Linda Cook, Howard Everson, Ida Lawrence, Mei Liu, Felicia Lyu, Gary Marco, Anne Marie Ninneman, Janet Schieferstein, Michael Walker, and Cathy Wendler improved the text of this paper, as did the suggestions of the editor and two anonymous reviewers, who insisted on knowing more of the whys and hows.

Contents

Figures

Abstract

The history of SAT® score scales is summarized, and the need for realigning SAT score scales is demonstrated. The process employed to produce the conversions that take scores from the original SAT scales to recentered scales in which reference group scores are centered near the midpoint of the score-reporting range is laid out. For the purposes of this paper, SAT verbal and SAT mathematical scores were placed on recentered scales, which have reporting ranges of 920 to 980, means of 950, and standard deviations of 11. (The 920-to-980 scale is used in this article to highlight the distinction between it and the old 200-to-800 scale. In actuality, recentered scores were reported on a 200-to-800 scale.) Recentering was accomplished via a linear transformation of normally distributed scores that were obtained from a continuized, smoothed frequency distribution of original SAT scores that were originally on augmented two-digit scales, i.e., discrete scores rounded to either 0 or 5 in the third decimal place. These discrete scores were obtained for all students in the *1990 Reference Group* using 35 different editions of the SAT taken between October 1988 and June 1990. The performance of this *1990 Reference Group* on the original and recentered scales is described. The effects of recentering on scores of individuals and the *1990 Reference Group* are also examined. Finally, recentering did not occur solely on the basis of its technical merit. Issues associated with converting recentering from a possibility into a reality are discussed.

I. Recentering and Realigning the SAT® Score Distributions

The choice of scales on which to report scores is one of a testing program's most fundamental and critical decisions. Scores are the most visible and widely used products of a testing program. The score is what the test-taker gets, and what score users use. The score scale provides the framework for the interpretation of scores. The choice of score scale has implications for test specifications, equating, and test reliability and validity, as well as for test interpretation.

Any test scale has one universal meaning that is shared by all, and a multitude of local meanings that are tied to specific local uses of a test. For a variety of reasons dealing with score interpretation and psychometrics, the original SAT scales were replaced in April 1995 by new recentered scales (Cook, 1994). The most salient reason for this change lies in the critical importance of the reference group to the universal meaning of score scales such as the SAT. The original SAT scales derived their universal meaning from a *1941 Reference Group* of slightly more than 10,000 test-takers. In this group, the expected SAT scores on the verbal section and the mathematical section were 500. Recentering replaced this *1941 Reference Group* with the *1990 Reference Group*.

We begin with a review of the various SAT scales that have been used since 1926. The properties of a well-aligned score scale are then described. The performance of this *1990 Reference Group* on the original scales is described. The process employed to produce the conversions that take scores from the original SAT scales to recentered scales, in which reference group scores are centered near the midpoint of the score-reporting range, is laid out. The effects of recentering on scores of individuals and the *1990 Reference Group* are also examined. We then discuss the effects of recentering on the interpretation of score differences among important subpopulations of the SAT test-taking population. Finally, limitations and generalizations follow a conclusions section.

II. Brief History of SAT Score Scales

The First 20 Sets of Score Scales

There have been over 20 different sets of scales used since the SAT exam's inception in 1926. This fact may shock most people who think of the scales established in 1941-42 as the original SAT scale. But a little logic would suggest that the 1941 or original scale as it is known (and will be called in this report) must have had at least one predecessor because the original ancestor of the current SAT exam was administered in 1926.

As can be inferred from Angoff and Donlon (1971) and Donlon and Livingston (1984), the SAT scales were, in essence, recentered every year from 1926 to 1939 as raw scores on the test were converted to scales scores with a mean of 500 and a standard deviation of 100 at every administration of the SAT. For economy of exposition, the numerical phrase {500/100} will be used as shorthand for setting the mean to 500 and standard deviation to 100. Until 1938, the SAT, like many tests in many nations today, was administered only once a year

and cross-year comparisons were of little interest. Thus, each June from 1926 until 1937, SAT raw scores were placed at the center of a 200 to 800 point scale with a mean of 500 and standard deviation of 100. Since the reference group changed from year to year, scores were not comparable across years.

In 1938, an important change occurred. The SAT was administered twice that year, in April and in June. The practice of setting a new scale within each year continued to occur. Scores in April were given a mean of 500 and a standard deviation of 100, as were scores in June. This practice only made sense if the April and June groups were equivalent in SAT math (SAT M) and equivalent in SAT verbal (SAT V). They weren't. The same practice was continued in 1939, as the 14th and 15th sets of SAT scales were established in April and May of that year.

By 1940, it was clear that setting scales anew with each administration was unfair to candidates who took the test with the more able cohort. So in 1940, the SAT V scored was scaled to {500/100} in April, and the June 1940 SAT V scores were linked to the April scale via common item equating. The April SAT M was also scaled to {500/100}. But the June SAT M was scaled to SAT V in June 1940. So the two SAT V administrations were on a common scale, while the two SAT M exams were not.[1] By the end of 1940, there had been 17 SAT M scales and 16 SAT V scales.

In 1941, the April SAT V exam was scaled anew again to {500/100}, and June was linked to April via common items, establishing a 17th SAT V scale. SAT M was scaled to {500/100} in April 1941, and the June M exam was scaled to the June SAT M, as had been done in 1941, producing the 18th and 19th SAT M scales.

In 1942, the SAT V was linked to the April 1941 exam through common items, as had been the June 1941 exam. Hence from 1941 on, the April 1941 scale served as the frame of reference for all SAT V scores until April 1995. This April 1941 scale, the 17th verbal scale, is the so-called original scale.

In 1942, the April SAT M was scaled to the April 1942 SAT V, which itself was linked to the April 1941 SAT V original scale, thereby establishing the 20th SAT M scale. In June of 1942, common item linked the June SAT M to the April 1942 SAT M, the first time that scores on a SAT M were equated. Thus, for the April 1942 SAT M, the 20th scale became the so-called original scale for SAT M, which was in place from 1942 until April 1995.

From 1943 until April 1995, these two original scales were in effect as new editions of SAT V were linked via score equating to April 1941, and new editions of SAT M were linked via score equating to April 1942. Since all tests from 1941 were either equated to the April 1941 SAT V scale or linked to it via a mix of concordances and equatings, in the case of SAT M in 1941 and 1942, the convention has been to call the April 1941-42 scales the original scales. Linkage to this set of original scales permitted comparisons of examinees over time, a practice that became more and more common as the SAT became more and more popular (Angoff and Donlon, 1971; Donlon and Livingston, 1984).

Growing Concerns About the Original (20th) Set of Scales

From 1941 until 1951-52, the SAT V mean dropped from 501 to 476, and the SAT M mean dropped from 502 to 494, such that the SAT V and SAT M means differed by 18 points in 1951-52. Ten years later, in 1961-62, the SAT V mean had dropped an additional two points to 474, while the SAT M mean increased by one point to 495.

By the late 1950s, concern with the 20th set of scales reached a state that required a series of special studies led by S.S. Wilks (1961). His report, *Scaling and Equating College Board Tests*, published in 1961, examined growing problems with the SAT scales that that been set in 1941-42. The report acknowledged that the educational arena had changed dramatically between 1941 and 1961, and that this change led to a major shift in the SAT test-taking population. The test-taking population was no longer mostly restricted to a selective self-selected group of students applying to Ivy League colleges and other prestigious Eastern colleges. World War II had changed the role of women. The GI Bill had expanded educational opportunity. College Board member colleges had gone from 44 to 350 between 1941 and 1961, nearly a nine-fold increase. Many of these new colleges came from the South and the West. Scholarship programs had also expanded opportunity. These increases in educational opportunity resulted in changed populations and presented scaling problems for the 1941-42 scales.

Despite these dramatic population shifts, the approximately 20-point differential between SAT V with a mean near 475 and SAT M with a mean near 495 that existed around 1960 was not big enough to warrant an adjustment of scale according to Wilks who recommended:

> The College Board should make no attempt to establish a new master reference population as a

[1]Linking the math test to the verbal test in June 1940 did not equate the April and June SAT M tests. Practitioners who are unaware of the fact that equating requires a common construct may also perform this type of cross-construct link. SAT verbal/math cross-construct linkings are clearly population dependent as has been shown by Dorans and Holland (2000) with male and female subpopulations. This type of cross-construct linking may be resorted to out of necessity when nothing else can be done.

> basis for interpreting either the present scale or a new one. Instead, the scale now in use should be continued, but with renewed determination to freeze it and make it as invariant as possible over time. In addition, an increased and continuing effort should be made to develop and publish normative information which will be of maximum use for the various groups of College Board users. (p. 3–4 of *Scaling and Equating College Board Tests*)

To paraphrase the recommendation, the 1941 reference population had lost its meaning by 1961, if not sooner. But then, any reference population will eventually lose its meaning. And in addition, changing reference populations will induce consternation among those who use scores and will be viewed with intolerance by the same. Therefore, continue to use the 1941 scale and use caveats to compensate for scale shortcomings. Try to avoid equating and scaling blunders in the future.

By the time of the Wilks report, it was clear that the SAT V and SAT M scales were no longer aligned, and that the disruption had occurred somewhere prior to 1952. Trying to figure out what happened prior to 1952 is speculative at best. In fact the Wirtz panel convened in the mid-1970s to investigate why SAT scores had declined dramatically after 1963 thought it unwise to consider data from before 1963,

> The panel has considered how far back to go in trying to analyze this scoring pattern. A 20-year comparison (1957–1977) would show the same decline that a comparison of 1963 to 1977 figures does….The statistical evidence for that earlier period is exceedingly thin, however, except for SAT scores themselves. We have accordingly concentrated on the 1963–1977 decline:...(page 5 of *On Further Examination*).

The real decline in SAT scores did not start until after the Wilks report was issued. Shortly after the Wilks report, from about 1963 until 1980, both SAT V and SAT M means dropped noticeably from about 475 for SAT V to around 425, and from about 500 to 470 for SAT M. Now the difference in SAT V and SAT M mean scores was close to 45 points. By 1990, the SAT M mean had increased to near 475, while the SAT V mean remained around 425, a 50-point difference.

Except for the famous score decline of the mid-1960s to late-1970s, SAT mean scores have been remarkably stable. Prior to this dramatic decline, mean SAT V scores for all test-takers on the 1941 scale ranged in the 470s from 1951-52 to 1965-66, while mean SAT M scores during that same time ranged from 490 to 502. From 1980 until 1995, mean scores on the 1941 scale for the College Bound Senior Cohort have ranged from 422 to 431 for SAT V and 466 to 482 for SAT M. Outside of the period of the decline studied by the score decline panel and reported in *On Further Examination*, SAT means have been remarkably stable.

The decline halted by 1980. By then there was a definite need to realign the verbal and math scales. The SAT V and SAT M averages were 50 points apart. And there was a clear need to repopulate the top end of the score scale, especially for SAT V.

Even before the Wilks report, the top portions of the SAT raw-to-scale were consistently characterized by large gaps between raw scores and scaled scores. New editions of the test, especially for SAT V, were not scaling out to 800. In other words, a perfect raw score would correspond to a 760 or 770 or 780. The score reporting policy was to award an 800 to a perfect raw score. Hence the top score would be an 800, but one omission out of 85 items might cost a student 30 to 40 points.

Although a number of palliatives were applied to this enduring problem, each had its drawbacks. At one extreme, there was an approach which spread out the problem across many score levels, achieving something that looked good at the expense of damaging the comparability of scores above 700. For example, to deal with an edition that scaled only to 770, this approach might:

> add 10 points to scores in the 690 to 720 range, converting them from 700 to 730;
>
> add 20 points to scores in the 720 to 750 range, converting them from 740 to 770;
>
> add 30 points to scores in the 750 to 770 range, converting them from 780 to 800.

Repeated application of this approach to SAT V editions degraded the comparability of scores above 700, an area of importance.

At the other extreme, there was an approach that tried to meet the top score = 800 requirement while degrading less of the score scale. This approach permitted 20-point gaps, and would convert the 770 to 800, and then add 20 points to the next score (changing 760 to 780) and 10 points to the next (changing 750 to760), leaving the rest of the scale alone. This approach, while less attractive cosmetically, would maintain the scale up through scores of about 750. All these palliatives did not address the source of the problem: By the end of the famous the score decline, the score scale had outlived its usefulness. The infrastructure for SAT scores needed major repair.

III. The Well-Aligned Score Scale

The utility of a score scale depends on how well it supports the inferences attached to its scores, and how well it facilitates meaningful interpretations and minimizes misinterpretations (Petersen, Kolen, and Hoover, 1989). The scale should be well aligned with the intended uses of the scores. For a test like the SAT, a broad range test for which high, middle, and low scores may be pertinent for an admissions decision, the degree to which the scale is well aligned depends on how the scale was originally defined and how well current score distributions fall on that scale. If scale alignment is desired for tests like the SAT, the well-aligned scale should possess seven properties.

First, the scores of the reference group used to define the scale should be *centered* near the midpoint of the scale. The average score (mean or median) in the reference group should be on or near the middle of the scale.

Second, the distribution of aligned scores for the scale-defining reference group should be *unimodal,* and that mode should be near the midpoint of the scale.

Third, the distribution should be nearly *symmetric* about the average score.

Fourth, the shape of the distribution should follow a commonly *recognized form*, such as the bell-shaped normal curve.

Fifth, the *working range* of scores should extend enough beyond the *reported range* of scores to permit shifts in population away from the scale midpoint without stressing the endpoints of the scale.

Sixth, the number of scale units should not exceed the number of raw score points, which is usually a simple function of the number of items. Otherwise, unjustified differentiation of examinees may occur.

Seventh, a score scale should be viewed as infrastructure that is likely to require repair. Corrective action should be taken whenever average score distributions of current populations move sufficiently far away from the midpoint, or when distributions move far enough away from one of the endpoints to jeopardize the integrity of the scale at that endpoint, or when reference groups lose their relevance.

The recentering of the SAT scales was guided by these seven desiderata.

The seventh property was invoked to argue that the SAT scale should not be fixed in stone, but be flexible enough to change to keep up with the changes in the population of test-takers. The reasons for the first four properties are self-evident. If you want to maximize the longevity of the scale, you *center* the score distributions at the *center* of the score scale. Most human attributes have *unimodal* distributions. Given the symmetric or nearly symmetric nature of so many distributions of attributes, it seems logical to start with a *symmetric* distribution. The normal distribution is a unimodal symmetric distribution with a mathematically compact form that has known properties. The fifth property allows the distribution of scores to shift over time before the highest actual score is lower than the maximum reported score, or before the lowest actual score is higher than the minimum reported score. When the highest actual score falls short of the maximum reported score, then scores at the top end of the scale may be forced up to the maximum reported score via a scale-stretching process that may not produce exchangeable scores across editions of the test. Scores may be misinterpreted. Like the first property, having the working range subsume the score reporting range allows a score scale to be useful longer. The sixth property is the fundamental requirement that there be at least one item for each scale score point.

The placement of a unimodal, symmetric score distribution at the center of a reported score scale that is broad enough to accommodate shifts in the distribution should ensure that score interpretations are consistent and meaningful for an extended period of time. Provided the population of examinees is fairly stable, as is often the case with large populations, the score scale should be able to bear the subtle and slow-moving shifts in score distributions associated with that stable population.

Note that these seven properties deal with the location of observed scores distributions on the reported score scale, and that no mention is made of unobservables. Since equated observed scores are reported, and it is their characteristics over time that are of primary interest, the focus of the first six desiderata used to recenter the SAT is on their distributional properties. Other approaches that involve classical test theory or item response theory could be used to reset scales, e.g., Kolen and Brennan (1995).

IV. Discrete Versions of the Original SAT Score Distributions

The 1990 Reference Group

The data employed for recentering were very close to, but not identical to, the data reported in the annual College-Bound Seniors National Report (The College Board, 1990). The group of 1,052,000 students whose

scores were used for defining the new SAT scales are referred to as the *1990 Reference Group*, as opposed to the 1,025,523 students who comprise the *1990 College-Bound Seniors Cohort*.

The *1990 Reference Group* included the most recent SAT V and SAT M scores of students who graduated in 1990 *and* who last took the SAT in either their junior or senior year of high school. In contrast, the *1990 College-Bound Seniors Cohort* included the most recent SAT V and SAT M scores of students who graduated in 1990 *and* who last took the SAT any time in high school through March of their senior year. Hence, the major distinctions between the *1990 Reference Group* and the *1990 College-Bound Seniors Cohort* were: (1) the inclusion of approximately 30,000 additional senior-year scores from the May and June administrations of the SAT (the second and fourth largest SAT administrations composed primarily of juniors); and (2) the exclusion of approximately 5,000 freshman- and sophomore-year scores. Because the SAT is designed primarily for juniors and seniors, the *1990 Reference Group*, as we have defined it, was a cleaner cohort for recentering score distributions. Because the *1990 Reference Group* contains scores of seniors who last took the test in either May or June of their senior year, the average SAT scores for this group were slightly lower (approximately 1.5 points) than for the *1990 College-Bound Seniors Cohort*. The 1990 cohort means were 424 for SAT V and 476 for SAT M. For the *1990 Reference Group*, these means were 422 and 475. The recentering process described herein was based on 1,052,000 scores that were culled from 35 editions of the SAT administered between October 1988 and June 1990.

Extrapolating Incomplete Raw-to-Scale Conversions

Equating procedures are used with each new form of the SAT to convert raw scores on each edition of the test to scores on the 200-to-800 scales. For SAT V, the raw scores were obtained by summing item formula scores across 85 items and rounding them to integers. For SAT M, the raw scores were the sum of 60 item formula scores. Item formula scores were obtained by assigning {1} to correct responses, {0} to nonresponses, and {-1/(k-1)} to incorrect responses, where k is the number of options on the item.

Ideally, there will be a raw score for every possible scaled score. In practice this may or may not occur. Incomplete conversions occur whenever bounded scaling functions are allowed to have different endpoints. For example, if a difficult form is equated back to an easy form, the highest score on the harder form may correspond to a score on the easier form for which there is no scale score. Crone and Feigenbaum (1992) developed an empirical approach for extrapolating incomplete raw-to-scale conversions. Their symmetric weighted mean/sigma procedure was applied to 23 different SAT V forms and 23 different SAT M forms of the 35 SAT V and SAT M forms employed in the recentering process. This procedure was employed to obtain unrounded estimated scaled scores for the highest and lowest raw scores. Note that these estimated scaled scores were *not* forced to scale out to 800. Instead they were allowed to go where the extrapolation procedure estimated that the scores would have scaled to if the conversion had not stopped abruptly.

Treatment of Highest and Lowest Scores

Approximately 1 percent of the *1990 Reference Group* had scores above 690 on the original SAT V scale and scores above 750 on the original SAT M scale. Anyone who answers all questions correctly on an SAT V or SAT M test automatically receives an 800 on the original SAT V or SAT M scale. At the other end of the SAT V score scale, approximately 1.5 percent of the students would have scored below 200 if 200 were not the minimum reported score.

For the purposes of the recentering process, the scores used were those that the students would have received if perfect scores were not set equal to 800 and scores below 200 were permitted. We did not want to carry over the effects of these score-reporting truncation and stretching practices into the recentered score distributions.

Discrete Score Distributions on the Original Scale

Figure 1 displays grouped frequency distributions of *1990 Reference Group* scores on the original SAT V and SAT M 200-to-800 scale. The midpoint of a 200-to-800 scale is 500. Percentages of scores that fall between 200 and 800 in 12 50-point intervals are displayed. Scores below 200 have been converted to 200. The percentage of scores reported for an interval, such as the 17.9 percent reported for the SAT V interval 400–440, is the number of scores in that interval divided by the total number of students.

Most of the scores are to the left of the 500 midpoint, particularly for the SAT V scale where 75 percent of the scores are below 500. For SAT M, 57 percent of the scores are below 500.

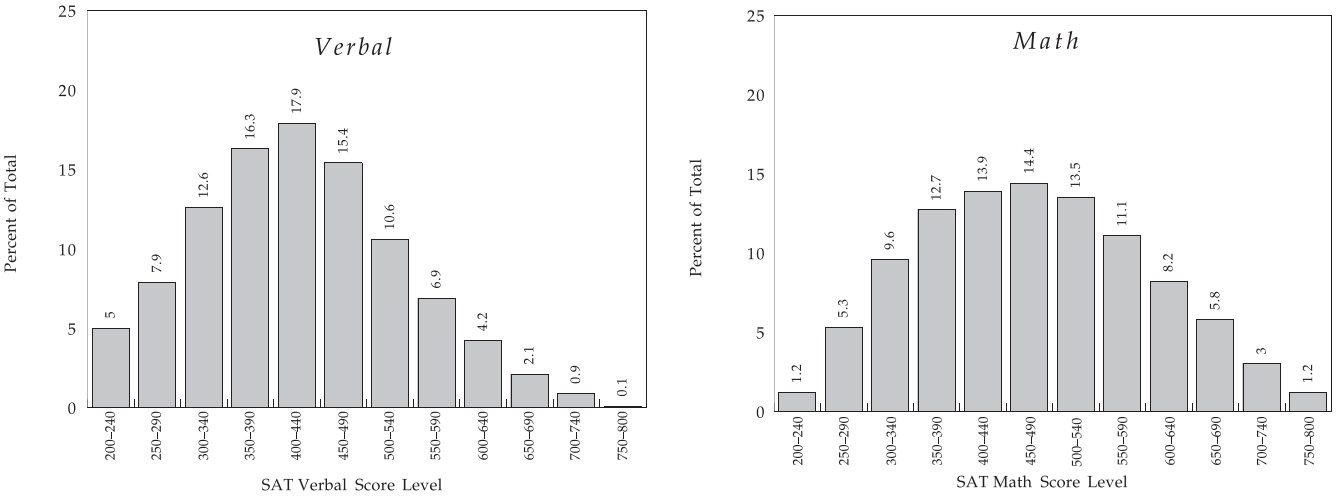

Figure 1. Distributions of SAT V and SAT M scores for the *1990 Reference Group* on the original scale.

For SAT V, the data can be grouped further into thirds of the scale:

Percentage scoring below 400 is	42 percent
Percentage scoring between 400 and 590 is	51 percent
Percentage scoring above 590 is	7 percent

For SAT M, the data can be grouped further into thirds of the scale:

Percentage scoring below 400 is	29 percent
Percentage scoring between 400 and 590 is	53 percent
Percentage scoring above 590 is	18 percent

The original scale for SAT V clearly was not aligned well with the score distributions for the existing SAT target population, college-bound juniors and seniors. The same problem existed for SAT M, but to a lesser extent.

V. Continuous Versions of the Original Score Distributions

Continuous versions of the distributions of scores on the original scale were needed in order to determine transformations that could map any original score onto a new recentered scale. Several steps were involved. First, the two-digit versions of the SAT scaled distributions were replaced with less discrete representations of the distributions of scaled scores. Then, these distributions were approximated by smoothed distributions. Finally, these smooth approximations to the observed distributions were made continuous.

Converting Formula Scores to Augmented Two-Digit Scales

Scores on the SAT V and SAT M tests are reported on a 200-to-800 scale. For about 30 years, the last digit of the three-digit score has been fixed at zero. Earlier, a full three-digit score was reported. The decision to fix the last digit at zero was made to discourage test-takers and score users from making arbitrary distinctions among students with virtually identical test scores, distinctions that could not be justified on the basis of the number of questions used to assess the students' mathematical or verbal proficiency. For example, a formula-score from a 60-item math test does not have the 601 pieces of information that a three-digit 200-to-800 scale implies. Thus fixing the last digit at zero effectively reduced the SAT scale to 61 points rather than 601.

For the recentering process, however, rounded two-digit scores on the 200-to-800 scale are too coarse for describing the scaled score distributions of 1,000,000+ examinees who took one of 35 editions of the SAT. After rejecting the two-digit version of the 200-to-800 scale because it rounded away too much information, formula scores were converted to three-digit scaled score, e.g., 201, 202, 203, …501, 502, …800, by applying the extrapolated formula score to unrounded scaled score conversions to each formula score. The conversions (mathematical and verbal) used depended on which edition of the test the examinee took. Figure 2 displays these three-digit scale score distributions for SAT V, while

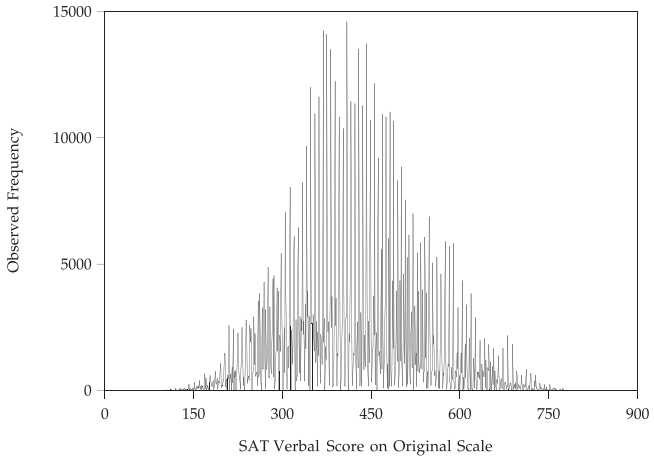

Figure 2. *1990 Reference Group* 3-digit score distribution (SAT V score on original scale).

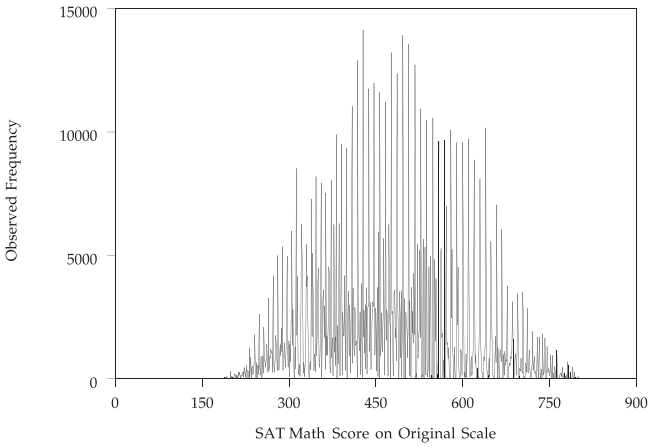

Figure 3. *1990 Reference Group* 3-digit score distribution (SAT M score on original scale).

the distributions for SAT M are depicted in Figure 3.

Distributions of these three-digit scaled scores are very irregular due to the fact that 35 distributions from administrations of variable volume were transformed onto the three-digit scale in their own unique ways. The tallest spikes belong to the three-digit scaled scores associated with the test editions administered at the largest volume administrations of the SAT. The smaller spikes belong to the three-digit scaled scores associated with the nonmajor administrations of the SAT. Any attempt at smoothing these spiked multigapped distributions was bound to fit the data poorly.

More regular score distributions were obtained when scores were placed on an augmented two-digit scale, i.e., rounded at the third digit to either 5 or 0 instead of ranging from 0 to 9 as in the three-digit case. In other words, possible scores were 200, 205, 210, 215, …790, 795, and 800. (Scores were actually allowed to extend below 200 and were not forced to 800.) Figures 4 and 5 depict these augmented two-digit scales for SAT V and SAT M, respectively. These distributions, though spiked, are at least suggestive of unimodal distributions that might be achieved through smoothing.

Approximating Augmented Two-Digit Scaled Score Distributions with a Smooth Function

We did not want to fit the spikes in the observed distribution, as they were a function of the spikes in rounded formula scores and where these spikes happened to map onto the 200-to-800 scale. Holland and Thayer's (1987) loglinear moment-matching smoothing procedure was used to produce a smooth approximation to the frequency distributions of scores on the augmented two-digit SAT V and SAT M scales. Several smoothings were computed, matching from 3 through 10 moments. For SAT V, there was

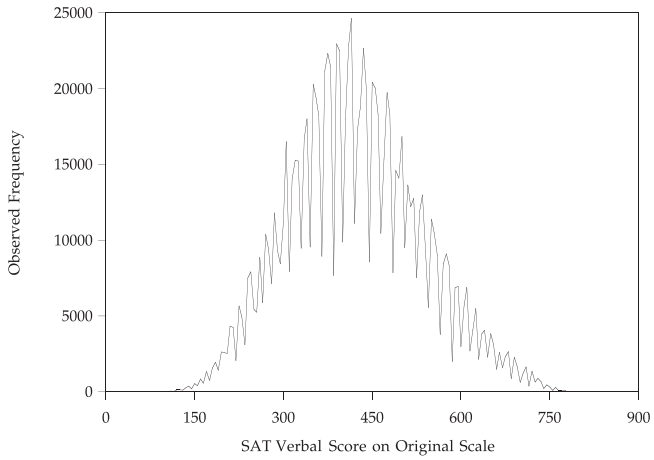

Figure 4. *1990 Reference Group* augmented 2-digit score distribution (SAT V score on original scale).

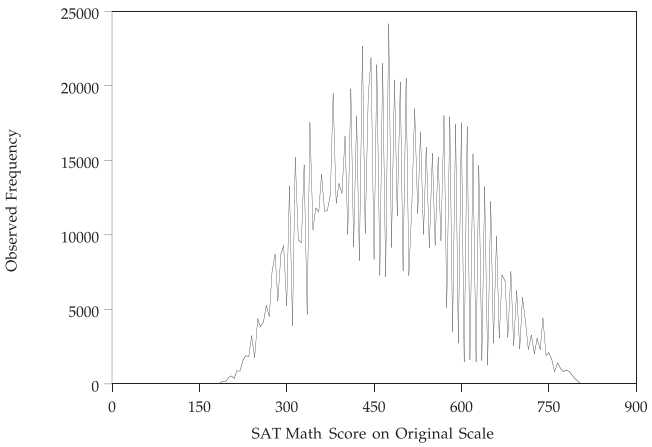

Figure 5. *1990 Reference Group* augmented 2-digit score distribution (SAT M score on original scale).

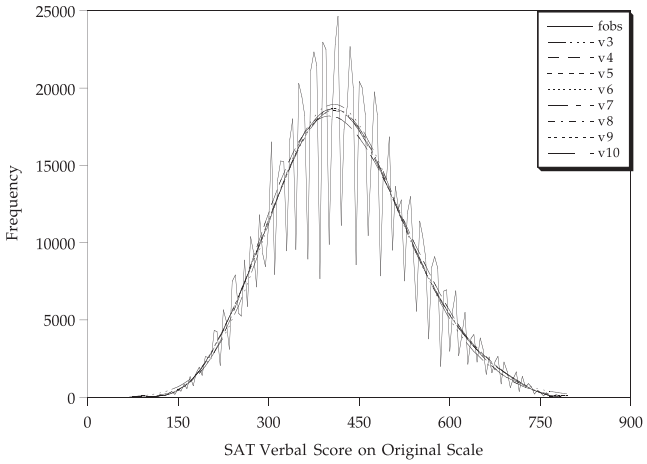

Figure 6. Observed and fitted (matching 3-to-10 moments) score distributions (SAT V score on original scale).

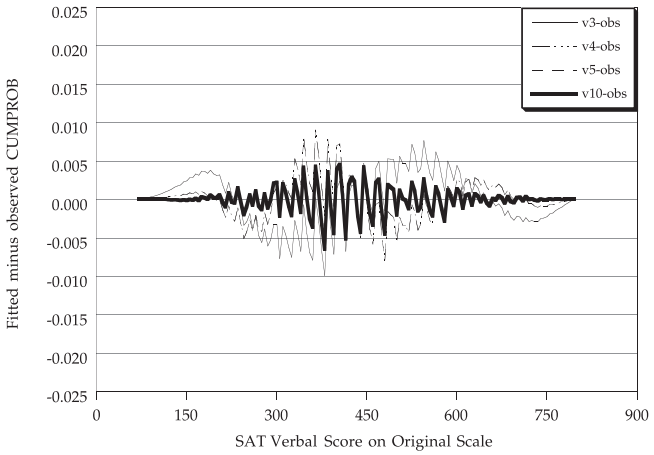

Figure 7. Cumulative probability fits for 3-, 4-, 5-, and 10-matched-moment solutions (SAT V score on original scale).

close agreement among all of these smoothed solutions, particularly for those that matched 5 or more moments, as can be seen in Figures 6 and 7. In Figure 6, the spiked observed frequency distribution and each of eight smoothed solutions are plotted. In Figure 7, residuals, in cumulative probability units, are plotted for the 3-, 4-, 5- and 10-moment solutions. The 3- and 4-moment solutions seem to be biased in opposite directions. The 3-moment solution does not fit well in the tails. In contrast, the 10-moment solutions seem to fit well, oscillating around the line of perfect fit, and rarely deviating from that line by .005. The oscillations are due to the undesired spikes in the observed data, and should not be misconstrued as undesirable fit. Given the large sample of over 1,000,000 examinees, we used the 10-moment solutions for both SAT V and SAT M because we had ample data to fit that many moments. Very weak smoothing was done to approximate the spiked distributions, just enough to remove the spikes in the observed distribution. In other words, 10 moments were matched instead of 5 or fewer to ensure that the smoothing followed the data more closely.

For SAT M, the various loglinear smoothings agreed less than they had for SAT V, as can be seen in Figures 8 and 9. In Figure 8, the spiked observed frequency distribution and each of eight smoothed solutions are plotted. In Figure 9, residuals, in cumulative probability units, are plotted for the 3-, 4-, 5-, and 10-moment solutions. Whereas the verbal score distribution was essentially normal, the math score distribution is clearly nonnormal. To achieve convergence among methods of smoothing for SAT M, more moments were required than for SAT V. In Figure 9, the poor fit of the 3-moment solution is quite striking. The 4- and 5-moment solutions are marked improvements over the 3-moment solution, yet visibly inferior to the 10-moment solution, which fits the data quite well. The 10-moment smoothing was selected for SAT M.

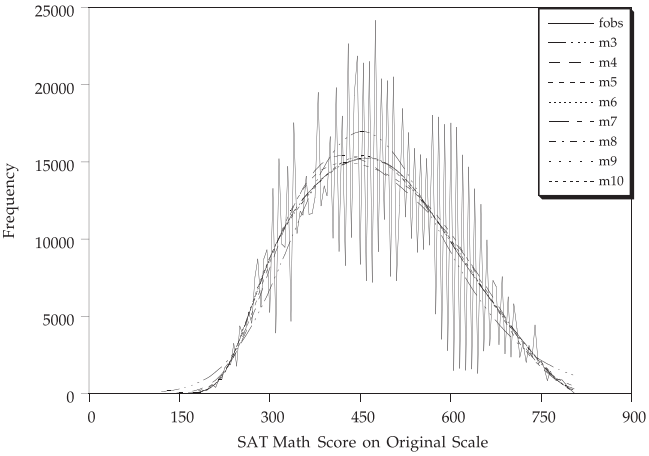

Figure 8. Observed and fitted (matching 3-to-10 moments) score distributions (SAT M score on original scale).

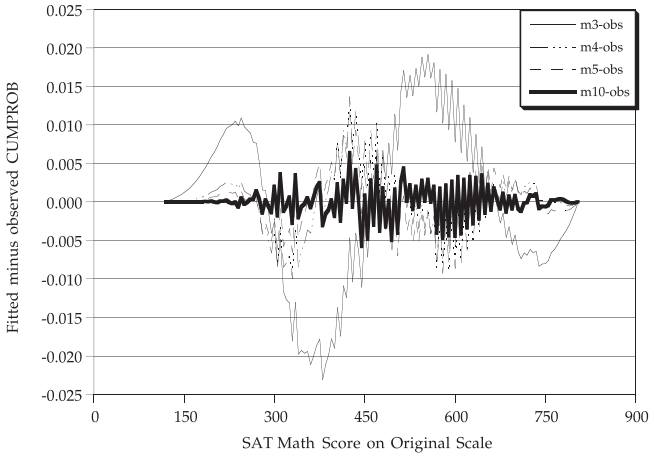

Figure 9. Cumulative probability fits for 3-, 4-, 5-, and 10-matched-moment solutions (SAT M score on original scale).

Continuization

The smoothed approximate score distributions for SAT V and SAT M were still discrete, i.e., values only existed for scores on the 200-to-800 scale in steps of 5, e.g., 200, 205, 210, …800. These discrete smoothed score distributions were then made continuous using the continuization step from the Holland and Thayer (1989) kernel method of score equating.

The kernel method is often thought of as a smoothing approach. In this context, it refers to a general class of functions for computing local averages according to different weighting functions. These kernel functions all possess a common set of properties (see Ramsay, 1991).

Using the kernel method to make a discrete distribution continuous can be thought of as spreading out the density at a discrete point onto an interval around that point. The Gaussian kernel function, which employs the well-known Gaussian distribution as the weighting function, was employed for recentering. Most of the weight was assigned to scores close to the evaluation point. This function tapers off gradually and assigns little weight to scores outside the bandwidth, h.

The tradeoff between bias and reduced variance influences choice of bandwidth, h. Larger values of h yield estimates based on larger sample sizes which reduces sampling variance. These larger values of h produce more smoothing, more bias, and less sampling variance. In contrast, smaller values of h involve less bias, but retain more sampling variance.

A bandwidth or continuization factor of $h = 3.69$ was employed for SAT V, while a continuization factor of $h = 3.51$ was employed for SAT M.[2] As can be seen in the residuals plots for Figures 10 (SAT V) and 11 (SAT M),

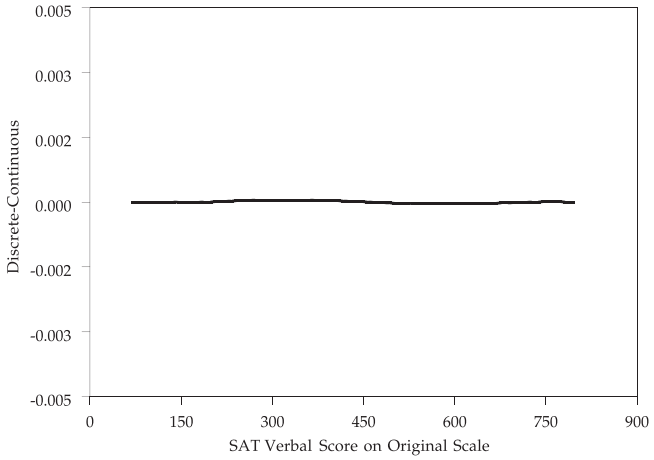

Figure 10. Fit of continuous function to smoothed discrete (10 moments matched) cumulative function for SAT V.

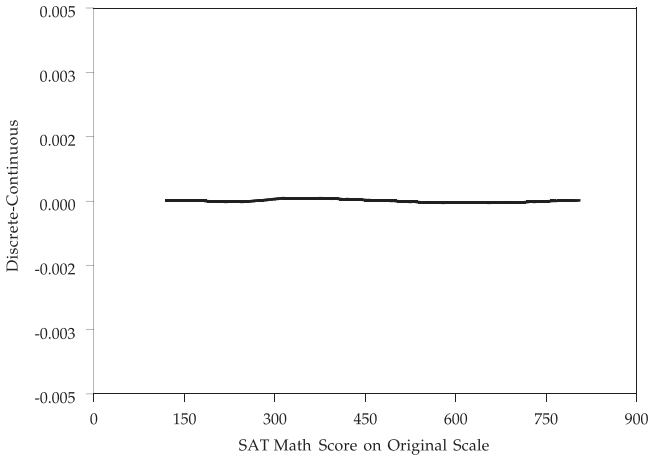

Figure 11. Fit of continuous function to smoothed discrete (10 moments matched) cumulative function for SAT M.

the continuous cumulative distribution functions fit the smoothed (10 matched moments) cumulative distribution functions very well. Continuization provides a continuous function that can be accessed at any possible scaled score to any number of significant digits to produce a relative frequency for that score.

VI. The Recentered and Aligned Scale

Normalization

The continuization function describing the smooth function that approximates the frequency distribution of SAT scores that was obtained from the observed distribution of scaled scores on the augmented two-digit scales was then normalized using the proportions to z-score transformation. This produced a function that converts any score on either the original SAT V scale or SAT M scale to a normalized score scale with a mean of 0 and a standard deviation of 1. The normalization process can be viewed as scaling the reference score distribution to a standard normal distribution using equipercentile methods.

Normalized scores were obtained via $z = \phi^{-1}[F(x)]$ where $F(x)$ is the continuous smoothed cumulative distribution of scores x on the augmented two-digit scale, ϕ^{-1} is the inverse of the standard normal distribution function, and z is the normalized score. This equation can be found in Kolen and Brennan (1995, equation 2.11, p. 36).

[2] Paul Holland and Dorothy Thayer, who also recommended 10 moments for smoothing, selected these values. These h values minimized the sum of squared differences between the height of the density at the score points times the interval size minus the height of the histogram for that score value.

This normalization step went a long way toward producing a scale that is well aligned because it produces a score distribution that is symmetric around its average score (mean, median, and mode of zero).[3] The next step was to take scores on this fundamental recentered scale and convert them to a scale that retained some characteristics of the SAT scale.

The Original 200-to-800 Scale and the 920-to-980 Scale

The essence of the original SAT V and SAT M scale is in the 61 points captured in the range of the first two digits of the three-digit scale. For purposes of this report, we will distinguish the recentered scales from the original scales by placing the recentered scales on a 920-to-980 metric.[4] The prefix 9 reminds us that the scale was established using data from the *1990 Reference Group*. We can also think of the 9 as a units marker for the new metric of these scores. Just as it is necessary to attach in., ft., or yd. to numbers that describe length, it makes sense to attach mnemonics to numbers that describe proficiencies on different scales. For this article, the prefix 9 will serve as our mnemonic unit marker. The working range of this 920-to-980 scale is still the familiar 20 to 80, the first two digits of the 200-to-800 scale.

Scores were transformed via a simple linear transformation from the unit normal scale to a scale with a mean of 950 and a standard deviation of 11. A standard deviation of 11 was selected over a standard deviation of 10, a more traditional choice, in order to avoid the scaling problems that have bothered the original SAT at both tails of the score distribution. With a standard deviation of 10, only one formula score per test is likely to convert to a 980. If a slightly easier test form is built, it would likely not scale out to 980 on the basis of empirical data. A standard deviation of 10 results in a scale in which the *working range*, the score range in which the equating is done, and the *reporting range*, the range of possible scores, are identical. One way of viewing the original SAT scales is to say the working ranges and reporting ranges are out of alignment, especially for SAT V, where the working range rarely exceeded 760, and the reported range was 200 to 800.

A standard deviation of 12 creates a working range that is too broad, in that 3 or more scores are likely to convert to a 980. In contrast, the 0-to-1 scores likely to convert to a 980 under a standard deviation of 10 are too few to compensate for the inevitable deviations from ideal specification that occur in practice. In contrast, a standard deviation of 11 is likely to have about 2 scores convert to 980. Hence a standard deviation of 11 was selected because it provided a working range that envelops the reporting range and which permits minor deviations from statistical specifications to occur that would not compromise the integrity of the score reporting scale.

In sum, SAT V and SAT M scores were placed on new recentered scales, which have reporting ranges of 920 to 980, means of 950, and standard deviations of 11. This was accomplished via a linear transformation of normally distributed scores that were obtained from a continuized, smoothed frequency distribution of original SAT scores that were on augmented two-digit scales, i.e., discrete scores rounded to either 0 or 5 in the third decimal place. These discrete scores were obtained for all students in the *1990 Reference Group* using 35 different editions of the SAT taken between October 1988 and June 1990. Conversions for these tests were extrapolated at the top and bottom when necessary, using the procedures described in Crone and Feigenbaum (1992). Scores were allowed to drop below 200 and were not required to scale to 800.

VII. Conversions from the Original Scales to the Recentered Scales

SAT V Conversion

Figure 12 displays the conversion from the original 200-to-800 scale to a recentered 920-to-980 scale for SAT V. For reference, the dashed line displays the conversion from the original 200-to-800 scale to the 920-to-980 scale obtained by simply dividing the original scale by 10 or dropping the inert trailing zero, and placing a 9 in front of the resulting numbers.

This recentering conversion is a monotonic transformation that never reverses the rank ordering of individuals in

[3] We chose to have the centered scores follow a normal distribution for two reasons: familiarity and symmetry. The normal distribution is symmetric and widely known. Symmetry ensures a centered distribution. If we had selected a less familiar symmetric distribution, we would have had to explain why we hadn't chosen the familiar normal distribution. So we chose the normal distribution for its symmetry and familiarity. We did *not* choose it because we believe that ability is normally distributed.

[4] In practice, scores on both the old and new scales have been reported on a 200-to-800 scale. When scores were first reported in 1995 on the new scale, an **R** was appended to them on score reports, and a footnote indicated when the change occurred. This practice was discontinued in fall of 2001. We use the 920-to-980 scale in this article to highlight the distinction between it and the old 200-to-800 scale.

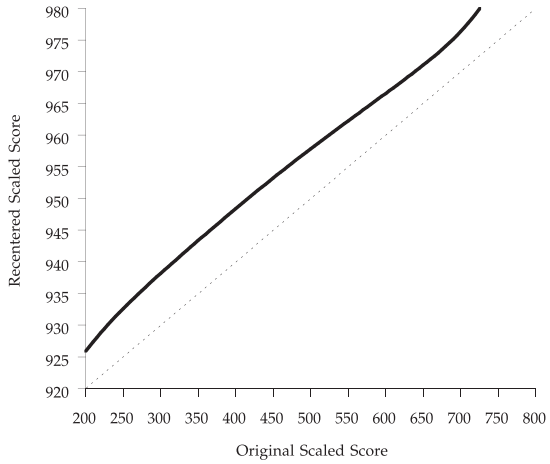

Figure 12. SAT V original to recentered scale conversion line.

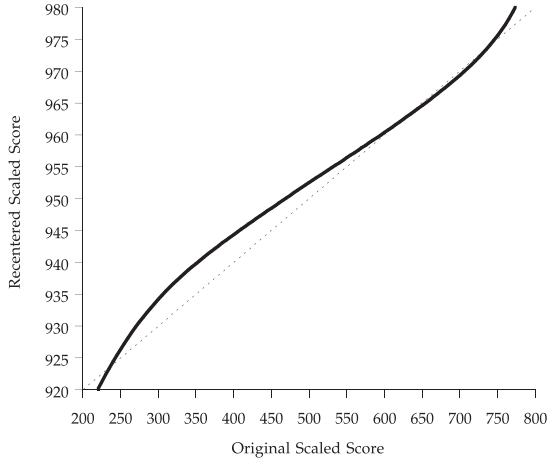

Figure 13. SAT M original to recentered scale conversion line.

a manner which allows ordering reversals to occur.[5] If the original score for person A is higher than the original score for person B, then the recentered score for person B will *not* be higher than the recentered score for person A.

The conversion from the original SAT V scale to the recentered SAT V scale is essentially linear. This means that most score levels are treated in essentially the same manner: Divide the original score by 10 and add about 7 to 8 points and then add 900 to arrive at the scale value on the recentered scale.

Scores that were below 200 on the original scale were rounded to 200; on the recentered scale, they are permitted to take on distinct values. For example, an original scale score of 190, which was reported as 200, is a 924 on the recentered scale.

At the top end of the scale, the large differences in reported scores associated with small differences in number of items answered correctly that occurred because of forcing scores to scale to 800 shrink on the recentered scale to a size that is more in line with a one-item=one-point rule. Hence, distinctions among top-ability students become more empirically based because scores at the top end of the recentered scale are more comparable across editions of the SAT than they were on the original SAT V scale.

SAT M Conversion

Figure 13 displays the conversion from the original 200-to-800 scale to a centered 920-to-980 scale for SAT M. Again, the dashed line displays the conversion obtained by dividing the original scale by 10 or dropping the inert trailing zero, and adding 900.

This recentering conversion is a monotonic transformation that never changes the ordering of individuals such that reversals occur: If the original score for person A is higher than the original score for person B, then the recentered score for person B will *not* be higher than the recentered score for person A.

The conversion from the original SAT M scale to the recentered SAT M scale is distinctly nonlinear. This means that different score levels are treated differentially. This is because of the nonnormal nature of the original SAT M distribution.

Scores below 240 and scores in the high 600s and low 700s convert to scale values that are lower than what would be obtained by dividing the original 200-to-800 scores by 10 and adding 900. Scores in the high 500s and low 600s are virtually unchanged (except for division by 10 and addition of 900), as are scores in the mid-700s. At the very top of the original scale, the solid black recentering conversion moves upward from the "divide-by-10 and add 900" reference line. The most dramatic effect in Figure 13 occurs between 250 and 550, where the recentering conversion increases scores over the reference line. This increase grows to a peak between 350 and 400 before it starts to decline.

Three linear sections can approximate the nonlinear nature of the conversion from the original 200-to-800 scale reasonably well:

1. If we consider division by 10 and adding 900 to be a "no change baseline," the recentering conversion converts scores below 240 to scale values that are *increasingly lower* than baseline values, while above 240, scores convert to scale values higher than baseline values, up to about 380;

[5] The imprecision inherent in rounding scores may lead to the differential breaking and creating of ties in reported scores, such that examinees with equal rounded scores on one scale may have different rounded scores on another scale. Likewise, examinees with different rounded scales on one scale may have the same rounded scores on the other scale.

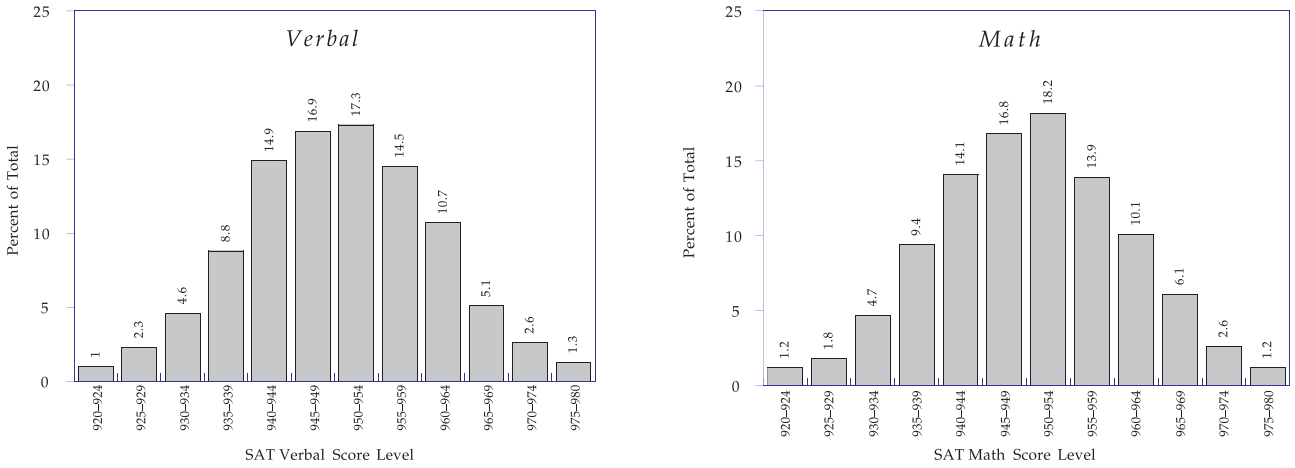

Figure 14. Distribution of SAT V and SAT M scores for the *1990 Reference Group* on the recentered scale.

2. Above 380, the conversion produces scale values, while higher than the baseline, that get closer and closer to the baseline until about 600;

3. Above 600, the conversion produces scale values that are very close to the baseline of "no change," until the very top of the scale where scale values will once again be higher than the "no change baseline."

Implications for Individual Scores

For SAT V, the conversion from the original scale to the recentered scale affects all scores in roughly the same manner. Hence, score *differences* between students at different score levels are virtually unchanged by recentering. The only exceptions to this statement are *reported* scores at either extreme of the score scale. Scores truncated at 200 are separated. Scores that were stretched out in the 700s are brought in line with each other, which leads to more comparability for SAT forms at the upper end of the scale. With the exception of scores at either end of the score distribution, score differentials are unchanged (except for division by 10).

For SAT M, score differentials are changed because of the nonlinear nature of the conversion. On the recentered scale, students with scores between 240 and 600 are closer to students with scores between 600 and 750 than they were on the original scale. At the ends of the distributions, scores below 240 are differentially lowered, while scores above 750 are differentially increased. These changes occur because, unlike the original SAT V scale, which is shifted down from its midpoint by 70 to 80 points, the SAT M score distribution on the original scale was asymmetric such that scores below 400 were relatively compressed, while scores between 400 and 700 were relatively more dispersed. The conversion in Figure 13 corrects this asymmetry and also centers the score distribution at the midpoint of the score scale.

Score Distributions on Recentered Scales

Figure 14 displays grouped frequency distributions of the *1990 Reference Group* scores on recentered SAT V and SAT M 920-to-980 scales. The midpoint of a 920-to-980 scale is 950. Percentages of scores that fall between 920 and 980 in 12 5-point intervals are displayed. In this and subsequent figures, scores below 920 have been converted to 920.

Approximately half the scores are to the left of the 950 midpoint and approximately half are to the right of the 950 midpoint for both SAT V and SAT M. (Any differences from exactly 50 percent on each side are due to the rounding inherent in grouped distributions and the fact that scores of 950 are included in the 950–954 interval.)

These recentered yardsticks for SAT V and SAT M are calibrated to match the score distributions for a more recent SAT target population.

A Comparison of the Original and Recentered Scales

Figure 15 combines the data from Figures 1 and 14 in a visual format that demonstrates how recentering yields better score balance within *and* across SAT V and SAT M. In the top portion of Figure 15 are the grouped frequency distributions for SAT V (on the left side) and SAT M (on the right side) for the original 200-to-800 scales. In the lower portion of Figure 15 are the grouped frequency distributions for SAT V (on the left side) and SAT M (on the right side) for the recentered 920-to-980

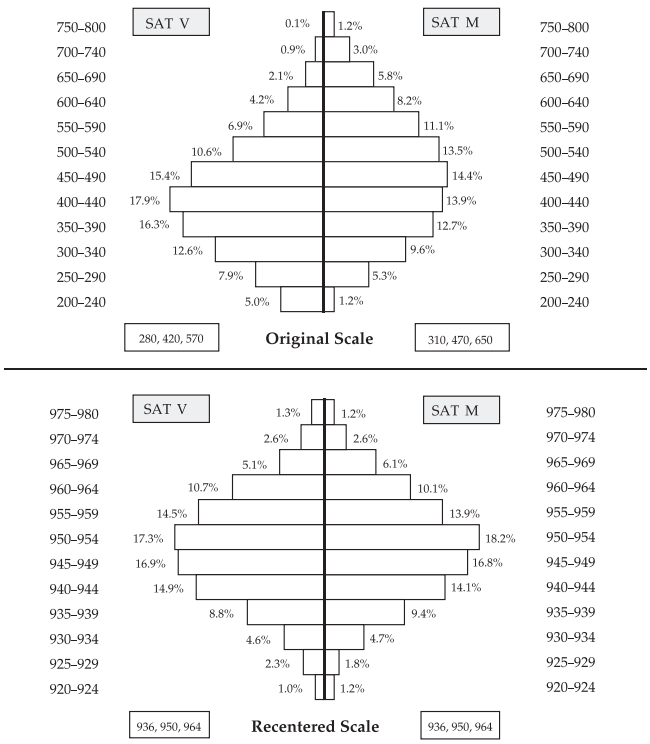

Figure 15. Distribution of SAT V and SAT M scores for the *1990 Reference Group* with 10%, 50%, 90% indicated.

scales. In addition to four boxes containing the SAT V and SAT M labels, there are four boxes that contain the scores corresponding to the 10th, 50th, and 90th percentiles on the two original and the two recentered scales.

On the original 200-to-800 scales (top portion), the off-center nature of the SAT V distribution is clearly not aligned with the SAT M distribution. The 50 percent point (or median) for SAT V is 420; its SAT M equivalent is 470. A 280 on SAT V is comparable to a 310 on SAT M, while a 570 on SAT V is comparable to 650 on SAT M. A 30-point difference at the 10th percentile becomes a 50-point difference at the 50th percentile, and an 80-point difference at the 90th percentile.

Figure 15 shows that the centered score distributions that were evident in Figure 14 for SAT V and SAT M are also very comparable across SAT V and SAT M. For both SAT V and SAT M on the recentered scale, the 50th percentile points (median) are 950, the 10th percentile points are 936, and the 90th percentile points are 964.

Scores on the original scale have meaning with respect to the 1941 group of 10,654 examinees. Scores on the recentered scales have meaning with respect to 1,052,000 more recent SAT test-takers. This is a very important point, and a major reason for recentering the SAT score distributions.

It is important to realize that while the recentering conversions did not change the ordering of individuals in the sense of how they are ordered by their scores on the edition of the test they took, they did affect how students are rank ordered when scores are compared across different editions of the SAT. Ideally, each SAT question should distinguish between students at one scale-score level from students at an adjacent level. Recentering made this ideal one-to-one relationship between number of correct answers (adjusted for guessing) and position on the score reporting scale more likely than it was with the original scales. The original scales had several many-to-one (e.g., a difference in three items answered correctly leading to a 10-point difference on the original SAT V scale) and one-to-many (e.g., two additional items answered correctly leading to a 60-point difference) conversions that occurred at the lower and upper portions, respectively. Replacement of these many-to-one clumps and one-to-many gaps that existed on the original scale improved the comparability of scores across editions of the test, and reduced the loss of information due to clumping.

VIII. Implications for Interpretations of Subgroup Performance

On the original SAT scales, the largest subgroup differences occurred among high scorers, while the smallest subgroup differences occurred among low scorers. This was because differences among individuals on the original SAT scale were largest for high scores and smallest for low scores. Subgroups are collections of individuals who reside in different proportions on different parts of the score scale. The effect of recentering on a particular subgroup depends on the effects of recentering on individuals and the mix of individuals who comprise that particular subgroup.

The particular transformations needed to align and center SAT score distributions suggest how these transformations might affect the relative performance of the various subgroups on the SAT. These transformations were presented in Figures 12 (SAT V) and 13 (SAT M). Whereas, the original scale was, in essence, a 20–80 scale with an inert trailing zero, the recentered scale, used in this report, is a 20–80 scale with an inert 9 preceding the 20–80.

For the SAT V scale, recentering does not have much of an impact on subgroup comparisons because the transformation is essentially linear through most of the score range, as seen in Figure 12.

On the SAT mathematical scale, the expected effects on subgroup differences are a function of the recentering process. As seen in Figure 13, on the original scale, scores below 400 were compressed and scores between 400 and 700 were stretched out. Recentering SAT M scores is expected to have an effect on subgroup comparisons, mainly because the standard deviation is reduced from 123 (on a 200 to 800 scale) to 110 (on a 200 to 800 scale). In particular, all groups are expected to appear closer to average on SAT M than they appeared on the original scale. Average scores for Asian American, white, and male groups are expected to appear less above average than they appeared on the original scale, while average scores for black, Hispanic, and female groups are expected to appear less below average than they appeared on the original scale.

Analyses were conducted in the *1990 Reference Group* with respect to gender (female and male students), and ethnicity (black, Hispanic American, Asian American and white students). Results are reported in separate sections for different groups.

Gender Comparisons

Figure 16 displays the effects of recentering on SAT V and SAT M score distributions for 547,474 female students. The format demonstrates how recentering yields better score balance within *and* across SAT V and SAT M, and is the same as the format for Figure 15. Figure 17 displays the effects of recentering on SAT V and SAT M score distributions for 504,526 male students.

On the original 200-to-800 scale (top portion of each figure), the highly off-centered nature of the SAT V distribution is quite evident. For female students, the 50th percentile (or median) for SAT V is 410, while for male students, it is 420. On the recentered 920-to-980 scale, the median score for both gender groups is 950, the midpoint of the scale.

On SAT M, the median for female students on the original 200-to-800 scale is 450, 40 points *higher* than the SAT V median of 410. On the recentered 920-to-980 scale, the median SAT M score for female students is 948, which is 2 (20) points *lower* than the SAT V median of 950.

On SAT M, the median for male students on the original 200-to-800 scale is 500, 80 points *higher* than the SAT V median of 420. On the recentered 920-to-980 scale, the median SAT M score for male students is 952, which is only 2 (20) points *higher* than the SAT V median of 950.

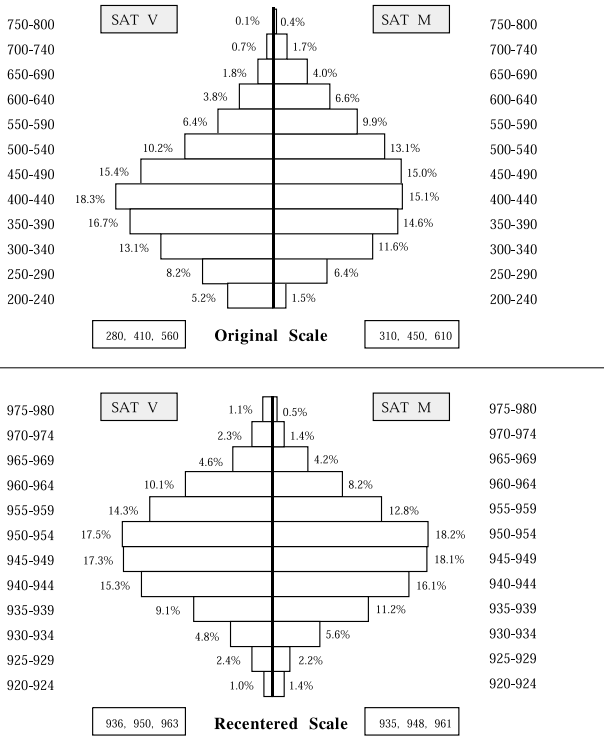

Figure 16. Distribution of SAT V and SAT M scores for the 1990 female reference group with 10%, 50%, 90% indicated.

Recentering brings male and female averages closer together, numerically, on SAT M, a mean difference of 3.8 (38) instead of 43, while leaving them virtually unchanged on SAT V, mean differences of 1 (10) and 10, as expected given the nature of the recentering conversions described earlier. Because the standard deviation of the *1990 Reference Group* on SAT M is 11 (110) on the recentered scale subgroup, differences are numerically smaller there than they were on the original scale on which the *1990 Reference Group* had a standard deviation of 123. SAT V subgroup differences are invariant to the scale shift because the standard deviation is essentially the same.

According to the original SAT scales, both female and male students appear markedly more able on SAT M than on SAT V, male and female students, on average, have similar verbal ability, and male students, on average, are noticeably more able mathematically than female students. According to the recentered scales, male and female students, on average, are still comparable verbally, and male students, on average, are still noticeably more able mathematically than female students. The recentered scales, however, indicate that female students, on average, are slightly more verbally able than mathematically able, while male students, on

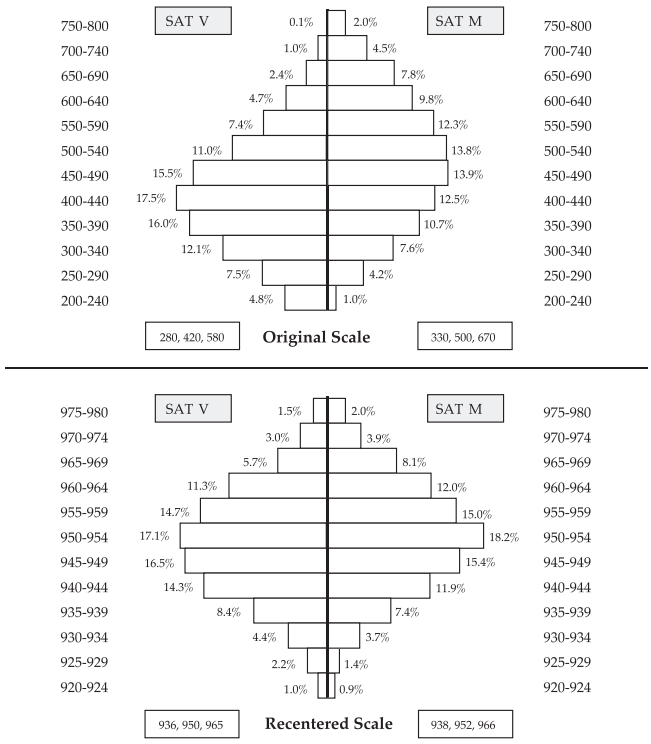

Figure 17. Distribution of SAT V and SAT M scores for the 1990 male reference group with 10%, 50%, 90% indicated.

average, are slightly more mathematically able than verbally able. Recentering produces score distributions for female and male students that are consistent with most well-known interpretations of gender performance data on mathematical and verbal tests.

The major effects of recentering for both gender groups was to realign SAT V scores and SAT M scores, place both sets of scores closer to the midpoint of the score reporting scale, and produce score distributions that are more consistent with well-known knowledge about gender differences.

While recentering realigned the SAT V and SAT M distributions, it did not alter the rank ordering of students within each score distribution. Table 1 displays the percentages of female and male students scoring above scores for SAT V (top panel) and SAT M (bottom panel) on the original and recentered scales that correspond to the top 1%, top 5%, top 10%, top 25%, top 50%, top 75%, top 90%, top 95%, and top 99% in the total group. The shaded center portion of this table contains parts of an equipercentile equivalence table between the original scale and the recentered scale, as

TABLE 1

Percentage of Scores in Both Gender Groups Above Certain Equivalent SAT V Scores (top panel) and Certain Equivalent Math SAT M Scores (bottom panel) on the Original and Recentered Scales

Verbal						
Female	*Male*	*Original*	*Total Group*	*Recentered*	*Female*	*Male*
% @ or >	*% @ or >*	*Scaled Score*	*% @ or >*	*Scaled Score*	*% @ or >*	*% @ or >*
1	1	690	1%	76	1	1
4	6	620	5%	68	4	6
9	12	570	10%	64	9	11
22	26	500	25%	57	24	28
47	51	420	50%	50	48	52
75	77	340	75%	42	76	77
89	90	280	90%	36	89	90
95	96	240	95%	32	95	95
99	99	190	99%	24	99	99
410	420	420	Median	50	50	50
417	427	422	Mean	50	49.5	50.5
110	114	112	SD	11	10.9	11.2
547474	504526	1052000	N	1052000	547474	504526
Math						
Female	*Male*	*Original*	*Total Group*	*Recentered*	*Female*	*Male*
% @ or >	*% @ or >*	*Scaled Score*	*% @ or >*	*Scaled Score*	*% @ or >*	*% @ or >*
0	2	750	1%	75	0	2
2	7	690	5%	68	3	8
6	13	650	10%	64	7	15
19	32	560	25%	57	20	33
43	57	470	50%	50	44	58
70	80	380	75%	43	69	79
89	93	310	90%	36	88	92
95	96	280	95%	32	94	96
99	99	240	99%	24	99	99
450	500	470	Median	50	48	52
454	497	475	Mean	50	48.2	52
116	126	123	SD	11	10.4	11.3
547474	504526	1052000	N	1052000	547474	504526

well as descriptive statistics for the reference group.[6]

If scores were not truncated at 200 on the original scale for SAT V, as many students (1%) would score below 190 as score above 690. This fact can be observed by comparing the first three columns of the first row (1%) with the first three columns of the ninth row (99%) in the upper panel of Table 1.

Comparing the Original Scale and Recentered Scale columns in this table reveals that the percentages of female and male students scoring above scores on the original and recentered scales that correspond to the top 1%, top 5%, top 10%, top 25%, top 50%, top 75%, top 90%, top 95%, and top 99% in the total group are *virtually unchanged* for SAT V and SAT M, as expected

[6] Because of rounding to integers, there are two inconsistencies in the SAT V and SAT M Recentered Scaled Score columns, at 1% (976 for SAT V and 975 for SAT M) and 75% (942 for SAT V and 943 for SAT M). Since rounding differences can occur throughout these tables, differences of 1 on the recentered scale or 10 on the original scale should not be overinterpreted.

with conversions that do not alter rank orderings of individuals.

Ethnic Comparisons

Figure 18 displays the effects of centering on SAT V and SAT M score distributions for 98,930 black students, while the effects of centering on SAT V and SAT M score distributions are displayed in Figure 19 for 63,624 Hispanic students, in Figure 20 for 73,754 Asian American students, and in Figure 21 for 708,310 white students.

Black Students. On the original 200-to-800 scale (top portion of Figure 18), the off-center SAT V distribution has a noticeable effect on the score distributions of black students where 90% of the scores are below 480, 50% are below 340, and 10% are below 230. On SAT M, the situation is only slightly better, where the 90th, 50th, and 10th percentiles are 520, 370, and 270.

On the recentered 920-to-980 scale (bottom portion of Figure 18), the median score for black students on both SAT V and SAT M is 942, much closer to the midpoint of the scale, and 10% of the black students score below 930 on both SAT V and SAT M, while 90% score below 956 on SAT V and below 954 on SAT M.

Black students are 30 points higher at the 50th percentile (or median) on SAT M than on SAT V on the original scale, whereas they have the same median (942) for both SAT M and SAT V on the recentered scale. Thus, the major effect of recentering for black students was to bring SAT V scores in line with SAT M scores and place both sets of scores closer to the midpoint of the score scale.

Hispanic Students. On the original 200-to-800 scale (top portion of Figure 19), the off-center SAT V distribution also has a noticeable effect on the score distributions of Hispanic students where 90% of the scores are below 520, 50% are below 370, and 10% are below 250. On SAT M, the situation is only slightly better, where the 90th, 50th and 10th percentiles are 580, 410, and 290.

On the recentered 920-to-980 scale (bottom portion of Figure 19), the median score for Hispanic students on both SAT V and SAT M is 945, much closer to the midpoint of the scale, and 10% of the Hispanic students score below 932 on SAT V and below 933 on SAT M, while 90% score below 959 on both SAT V and SAT M. The better balanced score distributions on the recentered scales are more comparable across SAT V and SAT M, and allow the test to make better distinctions among scores for the Hispanic students.

Hispanic students are 40 points higher at the median on SAT M than on SAT V on the original scale, whereas

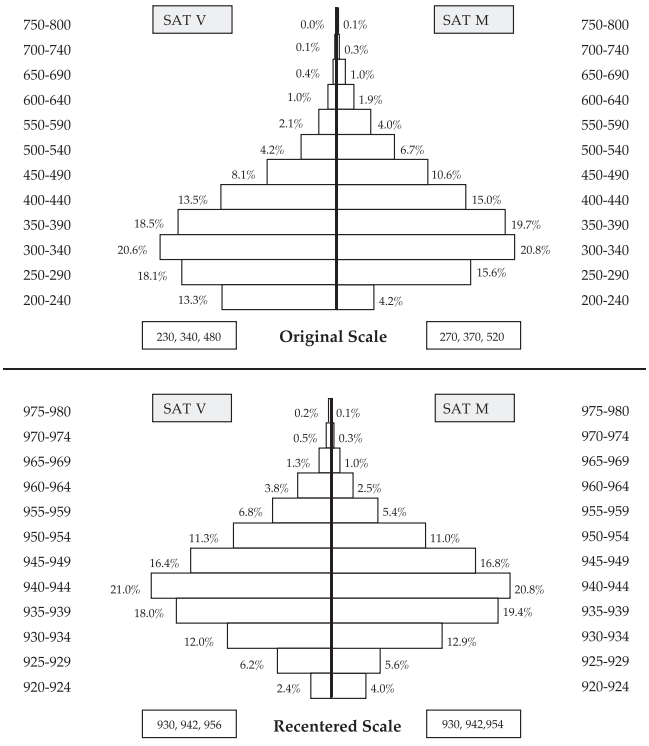

Figure 18. Distribution of SAT V and SAT M scores for the 1990 black reference group with 10%, 50%, 90% indicated.

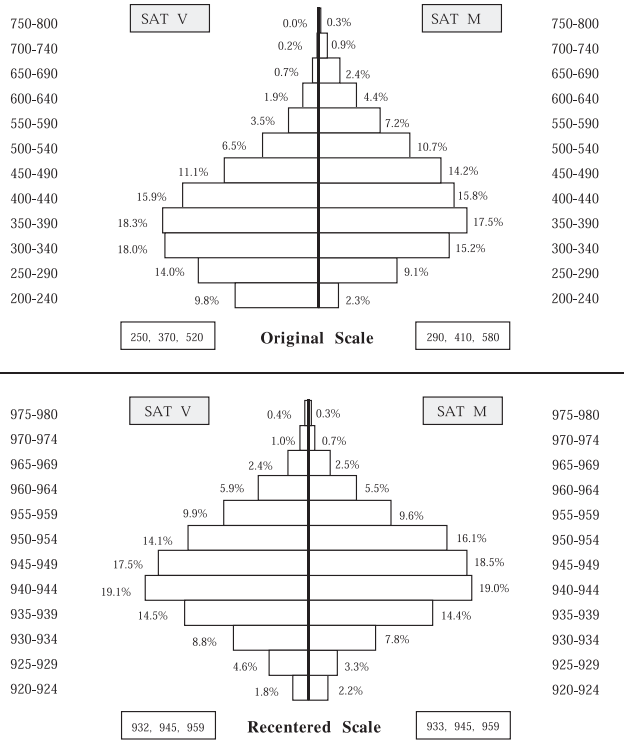

Figure 19. Distribution of SAT V and SAT M scores for the 1990 Hispanic reference group with 10%, 50%, 90% indicated.

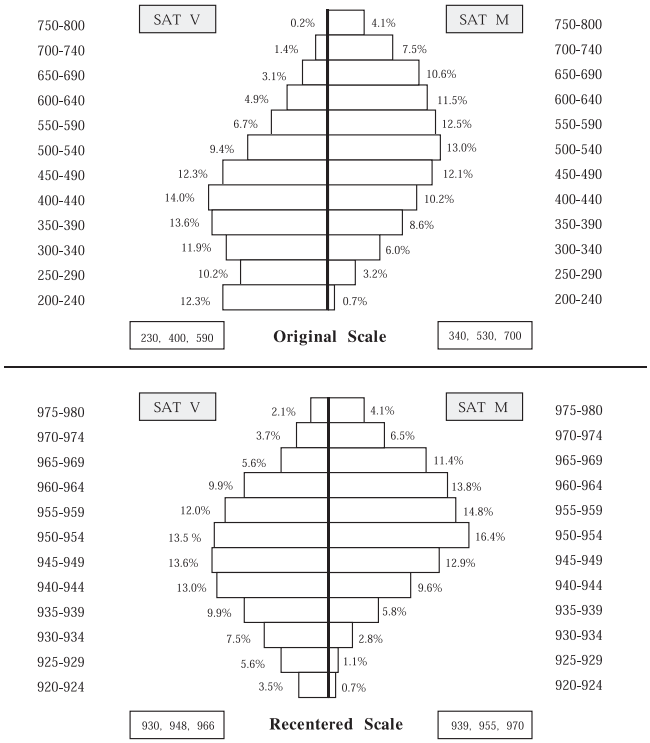

Figure 20. Distribution of SAT V and SAT M scores for the 1990 Asian American reference group with 10%, 50%, 90% indicated.

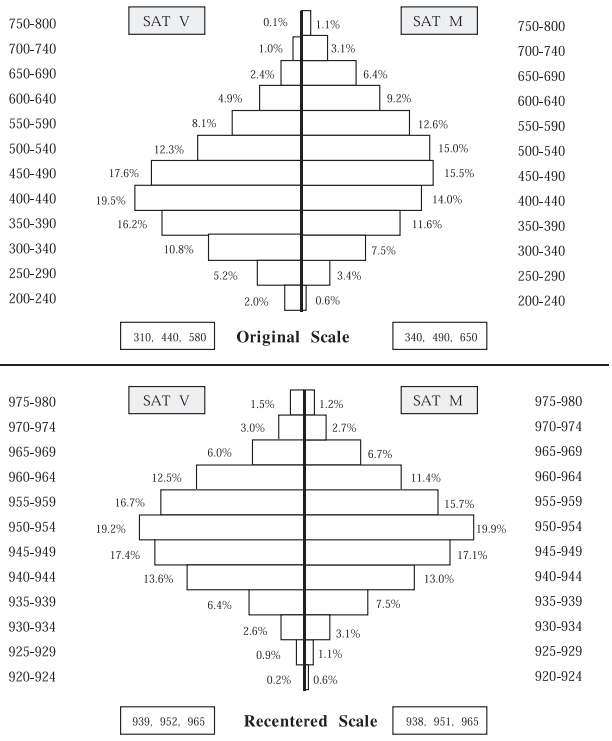

Figure 21. Distribution of SAT V and SAT M scores for the 1990 white reference group with 10%, 50%, 90% indicated.

they have the same median (945) for both SAT M and SAT V on the recentered scale. Thus, the major effect of recentering for Hispanic students was to bring SAT V scores in line with SAT M scores and place both sets of scores closer to the midpoint of the score reporting scale.

Asian American Students. On the original 200-to-800 scale (top portion of Figure 20), the off-centered nature of the SAT V distribution is very evident for Asian American students, where 90% of the scores are below 590, while 50% are below 400 and 10% are below 230. The peculiar nature of this distribution reflects the fact that it is a mixture of two distribution because of the sizable number of Asian American students who did not learn English as their first language. On SAT M, the situation is much better, where the 90%, 50%, and 10% points are 700, 530, and 340.

On the recentered 920-to-980 scale (bottom portion of Figure 20), the median score for Asian American students on SAT V is 948 and the median score on SAT M is 955. The SAT V median got 8 (80) points closer to the scale midpoint for Asian American students when the scales were recentered, while the SAT M median moved only 2 (20) more points farther away from the midpoint. Asian American students' verbal and mathematical proficiency seem much more balanced on the recentered scales, a difference of 7 (70) points, than the 130 point median differential on the original scales would have us believe.

The 90th percentiles of the Asian American students on the recentered scales are 966 on SAT V and 970 on SAT M, as opposed to 590 and 700 on the original scales. Recentering did not affect the top portion of the Asian American SAT M distribution, but it had a dramatic effect on the top and bottom of the SAT V distribution. Thus, the major effect of recentering for Asian American students was to bring SAT V scores more in line with SAT M scores and place their SAT V scores much closer to the midpoint of the score scale.

White Students. On the original 200-to-800 scales (top portion of Figure 21), the off-centered SAT V distribution is evident in the score distributions of white students, where 90% of the scores are below 580, 50% are below 440, and 10% are below 310. On SAT M, the situation is slightly better, where the 90th, 50th, and 10th percentiles are 650, 490, and 340, respectively.

White students are 50 points higher at median on SAT M than on SAT V on the original scale, whereas they are only 1 (10) point on SAT V (952) than SAT M (951) on the recentered scale (bottom portion of Figure 21). Thus, the major effect of recentering for white students was to bring SAT V scores in line with SAT M

scores and place both the SAT V and SAT M averages (medians and means) higher than the midpoint (950) of the score reporting scale.

Summary

The major effects of recentering for both gender groups was to realign SAT V scores and SAT M scores, place both sets of scores closer to the midpoint of the score reporting scale, and produce score distributions that are more consistent with current knowledge about gender differences.

The major effect of recentering for black students was to bring SAT V scores in line with SAT M scores and place both sets of scores closer to the midpoint of the score scale, which facilitates improved score interpretation.

The major effects of recentering for Hispanic students was to bring SAT V scores in line with SAT M scores and place both sets of scores closer to the midpoint of the score reporting scale. These changes facilitate improved interpretations of scores.

Recentering did not affect the top portion of the Asian American SAT M distribution, but it had a dramatic effect on the top and bottom of the SAT V distribution. Thus, the major effect of recentering for Asian American students was to bring SAT V scores more in line with SAT M scores, and place their SAT V scores much closer to the midpoint of the score scale.

The major effect of recentering for white students was to bring SAT V scores in line with SAT M scores and place both the SAT V and SAT M averages (medians and means) higher than the midpoint of the score reporting scale.

IX. Concluding Comments

The score scale provides the framework for the interpretation of scores. The choice of score scale has implications for test specifications, equating, and test reliability and validity, as well as for test interpretation. Realigning and recentering score distributions for the SAT I: Reasoning Test has had important implications for scores and their derivative products. It was not a step taken lightly.

Applied measurement does not occur in a vacuum of principles. It happens in actual settings where measurement principles compete with other realities. To effect significant change in applied measurement practices, we must listen carefully to what our customers say in order to help meet their needs. We must understand what is important to our clients, fashion solutions that are responsive to their needs, and use reasoned argument to persuade that the course of action we advocate will achieve its desired goals. To achieve progress, we must be able to compromise and adapt to practical realities.

Fixing infrastructure is necessary, often difficult work. In addition to identifying the problem and a solution, we have to explain how the solution solves the problem and then make sure the solution can be adapted into a larger framework of reality. There is always resistance to changing something with a well-established identity, especially something so visible and widely used as the SAT scale. The well-known College Board scale was widely recognized. Users of SAT scores had developed local meanings of the score scales over decades of experience. Recentering was essential work needed to improve the scale's properties and the quality of the inferences based on those properties. But it was disruptive work, like repairs to a heavily used thoroughfare inevitably are.

Although the resistance to change was fairly widespread and robust, score interpretation problems associated with the original scale necessitated a change. Prior to recentering, misconceptions about the old SAT scales abounded. Most notably, many thought the average score on the composite SAT V+M was 1000, when, in 1990, it was actually closer to 900. Many thought the scores were in effect centered on the existing scales because centered score distributions made sense to them, and that therefore the average student was below average. Many also thought that both the SAT V and SAT M scores had the same average, namely 500. These misconceptions about the old scales made it easier to argue that the scores ought to be recentered. Not even the most vigorous defenders of the old scale thought it would be a good idea to center scores on a 400-to-1600 scale 100 points below the midpoint of the scale, and to have the average verbal score be 50 points lower than the average math score.

Once a decision was made that centered score distributions were desirable, a dilemma had to be resolved. The trademark 200-to-800 score scale was a given; however, continued use of it could prove confusing. (The 920-to-980 scale is used in this paper to facilitate comparisons that would have been difficult to make had both sets of SAT scores been described side-by-side on the same 200-to-800 scale.) A scaling principle recommends that any significant change in the meaning of scores should be accompanied by scale redefinition to ensure that confusion is avoided when there is a change in scales.

To achieve the benefits of realigned SAT scales, compromise was necessary. This compromise was unpopular among some measurement colleagues who believed that the scale redefinition principle should never be compromised. As I saw it, the confusion would dissipate in

short order (as it has), while the benefits of scale realignment would linger. Fixing a roadway is confusing especially to those who follow a rote route, but once the road is fixed, the smooth ride is appreciated if only subliminally.

The College Board recognized that placing recentered score on the 200-to-800 scales could lead to confusion. They decided to add an **R** to the scores to distinguish it from the old scale. They knew they would be blamed for dumbing down the test (even though the difficulties of the Verbal and Mathematical tests were not changed). They knew many score users would object to the change. With input and assistance from ETS, they embarked on a massive information campaign that would inform the public of the recentering process.

What admissions staff had become accustomed to expect of their applicants, and what high schools expected in terms of their own performance, changed with recentering. Students, guidance counselors, and admissions officers were all provided with information that helped make the transition smooth. The universal meaning of the scores had changed with the shift in reference population from 1941 to 1990. Not only had the universal meaning of scores changed, but their many local meanings had changed also. Their information campaign attempted to make the transition from the old meaning to the new meanings as smooth as possible given the constraint that the numbers were unchanged.

In 1995, NCME gave the College Board its *Award for Outstanding Dissemination of Educational Measurement Concepts to the Public*[7] in recognition of the quality of this massive information campaign.

Despite this effort, there were some rough spots. In particular, those who had tracked longitudinal trends on the SAT were forced to rethink some of their premises. Theories that were not robust enough to withstand a scale change had to be discarded. And the phrase "dumbing down the test" was heard with disheartening frequency.

Which scales, the more than 50-year-old original scales or the recentered scales, represent "truth"? Neither. To believe that one set of scales represents "truth" is to reify the score scales in a way that scores on general intelligence tests were reified during the first half of the twentieth century (Gould, 1981).

Instead of asking which set of scales represents truth, we should ask which scales are more useful for various purposes. The 50-year-old original scales provided continuity with the past by referring students to a reference group of 10,654 students who took a test in April 1941, a reference group that had been maintained for over 50 years. The reference group had become outdated well before the Wilks (1961) report was issued. At the time of the Wilks report, the verbal and math scales were misaligned, but their means did not deviate markedly from each other or the midpoint of 500. By the end of the score decline, however, the scales were in need of repair. The recentered scales refer each student to a more recent cohort of 1,052,000 students who took the SAT and were likely to have graduated in 1990. These 1,000,000+ examinees define verbal and math scales that remain better yardsticks for today's more heterogeneous student population than did original scales which were rooted in a highly self-selected group of students. The recentered scales yield interpretations that are consistent with the percentile information routinely reported on score reports, and have better distributional properties than percentile scales.

The transformations from the misaligned and off-centered score distributions on the 1941 scale to recentered and realigned score distributions did not alter the rank ordering of scores on SAT V and SAT M. As a consequence, the percentages of students in all subgroups scoring above equivalent scores on the original and centered scales were virtually unchanged for SAT V and SAT M. Perceptions about individuals, and about gender, ethnic, and language subgroups were altered dramatically, however, by the recentering and realigning of SAT V and SAT M scores. Converting from the original scales to recentered SAT scales led to altered perceptions about the relative academic strengths and weaknesses of various subgroups, and individuals who comprise the SAT I testing population. These clearer perceptions by and about today's youth led to clearer perceptions of their preparedness for further academic work than would have been possible with the old scale.

The transition is now complete. The **R** was dropped in 2001. The college graduates in the year 2000 were the first complete cohort to have scores on the recentered SAT I scales. These students went through their last years in high school and their years in college with a markedly different perspective of their own abilities, particularly their verbal skills, than earlier generations had of their abilities. For decades, scores on the old SAT scales told us that both males and females were better in mathematical rea-

[7] Among the various recentering products developed by the College Board was a booklet containing 13 tables relating old and recentered scores on the SAT I (SAT), SAT II (Achievement Tests), and vice versa. (Prior to 1993-94, the College Board offered the Admissions Testing Program which consisted of the Scholastic Aptitude Test [SAT] and a series of Achievement Tests. These were replaced by the SAT I: Reasoning Test and the SAT II: Subject Tests, respectively.) Different tables are included that relate means on the new SAT I verbal and math scores to means and standard deviations on the old SAT scales, and vice versa. These mean conversion tables are needed because the score-to-score transformations are nonlinear and hence inappropriate for transforming means.

soning than in verbal reasoning. The new scales tell us that females are better verbally than they are in mathematical reasoning. The scale shift altered the way generations of students view themselves.

Will the SAT I scales need to be realigned? Will recentering be needed somewhere in the future? The reference group for the new SAT I scales is at least 10 years old. It is the only cohort for which recentered scores are perfectly centered and aligned. Its predecessor was in place for over 50 years. During those 50 years, the population grew dramatically and was described by a score distribution that had shifted towards one end of the scale, differentially for Math and Verbal.

Change has occurred since 1990, but not enough to warrant discarding the *1990 Reference Group*. For one thing, the *working scales* still contain the *reported scales*; all portions of the 200-to-800 scale are being used for SAT I. In addition, the trailing zero on the SAT I scale can give the impression of large differences. A 505 is only .5 scale points (on a 61-point scale) above the midpoint of the scale, 500; likewise, a 515 is 1.5 scale points above its midpoint of 500. Recent cohort means are in these neighborhoods.

While the scales are in very good working order, they should be monitored. Change happens. Changes in testing technology, such as the growing use of graphing calculators with CAS capabilities, might push score distributions away from the center, which would suggest a change in construct being measured or a breakdown in the equating model. Verbal and Math distributions may be pushed further apart by the changing composition of SAT test-takers as more nonnative speakers of English are tested. At some point in the future, the *1990 Reference Group* will become dated. In fact, educational reformers of all orientations would be pleased if their brand of reform produces dramatic lasting gains in performance of the SAT cohort on the SAT I and other exams. If that happened, the *1990 Reference Group* would need to be replaced and the scores would need to be centered and aligned again. The point is that the scale should be useful and support inferences based on it. The new SAT I scales are more useful for today's student cohort than the old scales would have been.

X. Generalizations and Limitations

Generalizations

The approach described herein can be applied to tests that are like the SAT I, i.e., broad range tests for which high, middle, and low scores may be pertinent for an admissions decision. The score scales for these tests should be well aligned with the intended uses of the scores. A well-aligned scale for tests like the SAT should possess the seven properties described in *The Well-Aligned Score Scale* section.

The details in this paper describe one approach for developing a well-aligned score scale. Such an approach could be used with tests like ACT, GRE, GMAT, TOEFL, and other broad range admissions tests. Other approaches can be employed as well. For example, Kolen, Hanson, and Brennan (1992) used the conditional standard error of measurement as the cornerstone of an approach that has been used with ACT.

Limitations

Placement exams, such as those of the Advanced Placement Program® (AP®), and certification tests need different kinds of score scales that have score interpretations tied to performance on a criterion, be it classroom performance in an introductory college course, or performance on the job. While the particular approach described herein is not appropriate for these exams, the appeal to principles is relevant. Any score scale needs to be defined according to a set of principled desiderata. And that set of principles needs to include guidelines for review and revision of the score scale.

This paper took a detailed look at the scales of one of the most visible, widely monitored exams in the world. Because the SAT I is so heavily used, it was important to realign its scales with their intended purposes. The concomitant disruption that accompanied recentering was analogous to that seen with the repair of a major thoroughfare. The disruption was a nuisance while it occurred, led to noticeable improvement after its cessation, and was forgotten as the improvement came to be taken for granted.

All score scales should be examined to see if they serve their intended purposes. Despite the extensive informational efforts conducted by the College Board and ETS, recentering was disruptive. There was a nontrivial cost associated with improved score interpretation for students, parents, colleges, and high schools.

Scale changes should not be made lightly, but they should be made when necessary.

References

Angoff, W.A. & Donlon, T. F. (1971). The Scholastic Aptitude Test. In W. A. Angoff (Ed.) *The College Board Admissions Testing Program: A technical report on research and development activities relating to the Scholastic Aptitude Test and Achievement Tests*. (pp. 15–45). New York: The College Entrance Examination Board.

The College Board (1990). *The 1990 profile of SAT and Achievement test-takers: College-bound seniors national report*. New York: The College Entrance Examination Board.

College Board (1977). *On further examination: report of the Advisory Panel on the Scholastic Aptitude score decline*. W. Wirtz, chairman. New York: The College Entrance Examination Board.

Cook, L.L. (1994, April). *Recentering the SAT score scale: An overview and policy considerations*. Paper presented at the annual meeting of the National Council on Measurement in Education. New Orleans, LA.

Crone, C.R. & Feigenbaum, M.D. (1992). *The shrinking tops and bottoms problem of the Scholastic Aptitude Test*. (SR-92-40). Princeton, NJ: Educational Testing Service.

Donlon, T. F. & Livingston, S.A. (1984). Psychometric methods used in the Admissions Testing Program. In T.F. Donlon (Ed.) *The College Board handbook for the Scholastic Aptitude Test and Achievement Tests*. New York: The College Entrance Examination Board.

Dorans, N. J. & Holland, P.W. (2000). Population invariance and the equatability of tests: Basic theory and the linear case. *Journal of Educational Measurement. 37(4)*, 281–306.

Gould, S.J. (1981). *The mismeasure of man*. New York: WW. Norton.

Kolen, M. J., Hanson, B.A. & Brennan, R.L. (1992). Conditional standard errors of measurement for score scales. *Journal of Educational Measurement, 29*, 285–307.

Kolen, M.J. & Brennan, R.L. (1995). *Test equating: Methods and practices*. New York: Springer-Verlag.

Holland, P.W. & Thayer, D.R. (1987). *Notes on the use of log-linear models for fitting discrete probability distributions* (RR-87-31). Princeton, NJ: Educational Testing Service.

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

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

Ramsay, J.O. (1991). Kernel smoothing approaches to nonparametric item characteristic curve problems. *Psychometrika, 56*, 611–630.

Wilks, S.S. (Ed.) (1961). *Scaling and equating College Board tests*. Princeton, NJ: Educational Testing Service.