Research Report

# Using Score Equity Assessment to Evaluate the Equatability of the Hardest Half of a Test to the Total Test

Neil J. Dorans

Karin Zeller

Research & Development

October 2004
RR-04-43

# Using Score Equity Assessment to Evaluate the Equatability of the Hardest Half of a Test to the Total Test

Neil J. Dorans and Karin Zeller

ETS, Princeton, NJ

October 2004

**Abstract**

In an article published in the spring 2003 issue of *Harvard Educational Review*, Roy Freedle stated that the SAT® is both culturally and statistically biased. Freedle proposed a solution to this bias, which involved using a half-test made up of the most difficult items culled from complete SAT examination. His claims, which garnered national attention, were based on serious errors in his analysis. In Dorans and Zeller (2004), we demonstrated that the effects Freedle reported are reduced substantially when the data are analyzed correctly. Here, we describe a sound way of assessing whether the current SAT scoring procedure is fair, and we examine what happens when we subject the Freedle's hard half-test to a score-equity assessment.

Key words: DIF analysis, differential score prediction, SAT, score bias, score equity assessment

**Acknowledgments**

The authors are grateful to several colleagues for their reviews and comments on earlier versions of this work, especially Shelby Haberman and Michael Walker. Dan Eignor, Jinghua Liu, and Ed Kulick were helpful with this particular paper. The ideas and opinions expressed in this report are those of the authors and do not represent the opinions of the ETS.

## Overview

In an article published in the spring 2003 issue of *Harvard Educational Review*, Roy Freedle proposed a new scoring method for the SAT® that he believed would correct for what he perceived was existing ethnic bias in the test. In particular, in the first line of his abstract, Freedle (2003) states, "The SAT has been shown to be both culturally and statistically biased against Blacks, Hispanic Americans and Asian Americans." Freedle's claims garnered national attention via an article that appeared in the *Atlantic Monthly* (Mathews, 2003). His claims, however, are based on serious errors in his calculations (Dorans, 2004a).

Freedle recommended using two scores for reporting the performance of Black examinees on the SAT-Verbal: one based on questions that comprised the hardest half of the SAT-Verbal exam (called SAT-R) and one score based on all the items. What would be reported would then have two components, one common across all subgroups (the regular SAT-V score), and one subgroup-specific aspect (the R-score). White examinees would receive a score based on only the full set of items. There are a number of issues associated with the R-score approach. First, as noted, R-scores would be based only on the hardest half of the test. Second, R-scores are created using performance data from Black examinees. That is, R-scores, as proposed by Freedle, employ a subgroup-specific scoring procedure to place the R-score on a 200-to-800 scale. In others words, the same level of performance on the hard half-test would yield different scores on a 200-to-800 scale, depending upon which subgroup the examinee belongs. Finally, it follows from the logic of Freedle's proposed scoring procedure, if extended to the extreme, that it is possible for a Black female to have three different R-scores: one for being Black, one for being female, and one for being a Black female.

Elsewhere we have addressed several flaws in Freedle's analysis (Dorans & Zeller, 2004). Here we describe a sound way of assessing whether the current SAT scoring procedure is fair and examine what happens when we subject Freedle's hard half-test score to a score-equity assessment (Dorans, 2004b). In contrast to these earlier papers, which focused on the flaws in Freedle's analysis, this paper presents an approach that can be adopted to determine whether test scores are equatable across subgroups.

**Assessing the Fairness of Score Conversions**

Assessing tests for fairness is difficult. Fairness is a word that engenders passion. Often, readers seem to be blind to flawed analyses such as Freedle's because the flawed analysis tells them something they might want to hear. Fortunately, there are better ways to assess fairness than the flawed differential item functioning (DIF) analyses and score-linking analyses employed by Freedle.

Dorans (2004b) introduced a procedure for fairness assessment that complements two existing procedures and directly addresses the question of whether or not scores on different test editions are being fairly linked. The two existing procedures are DIF analysis (Holland & Wainer, 1993) and differential score prediction. DIF focuses on whether each item in a test measures the construct of interest in the same way across different subgroups. Differential score prediction examines whether a test score in conjunction with other information predicts performance on some criterion (e.g., grades in college) in much the same way across different subgroups.

The newly introduced third procedure, score equity assessment (SEA), asks the following question: Does the test score measure what it measures in the same way across different subpopulations as it does for the full population? All three procedures share a common form or structure in that all three involve checking for invariant relationships across subpopulations. However, the three differ with respect to function. DIF is a test-construction tool that can generate interesting hypotheses for further study, preferably in the ways suggested by Schmitt, Holland, and Dorans (1993). Differential prediction is an important tool for evaluating fair use of scores. SEA assesses whether the final products of the test-construction process can be used interchangeably and in the same way across subgroups. Perhaps Freedle thought he was examining this score-equity issue in his analysis.

Is the linking relationship between two hard Freedle half-tests the same across different groups; that is, the Total group, the Black group and the White group? Figure 1 depicts linking a hard half-test or easy half-test and a total test to another total test.

In Figure 1, T represents the total test (of which there are two, $T_1$ and $T_2$), which is an entire SAT-Verbal exam comprising 78 questions. H represents the hard half-test, which is composed of the 39 hardest questions from that SAT-Verbal exam, and E is the easy half- test, composed of the 39 easiest questions from that same exam. There are two hard half-tests ($H_1$ and

$H_2$) and two easy half-tests ($E_1$ and $E_2$). The dotted lines running from right to left indicate that the two tests were directly linked via an equivalent-groups equating. (See Kolen & Brennan, 2004, or von Davier, Holland, & Thayer, 2004, for a discussion of equating designs.) In an equivalent-groups design, the tests to be linked are administered to separate groups of examinees that are assumed to be equivalent. $E_2$ is linked to $E_1$ in this fashion, as is $T_2$ to $T_1$, and $H_2$ to $H_1$.

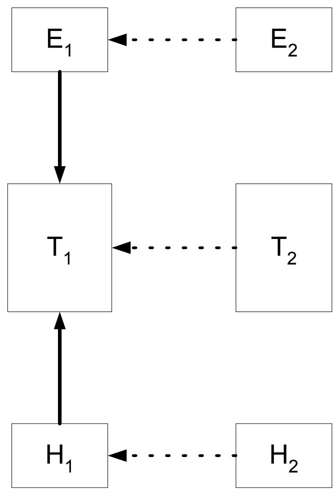


***Figure 1*. A linking schematic.**

The solid lines extending from $E_1$ and $H_1$ to $T_1$ on the left-hand side indicate a single-group linking in which the same examinees took both tests. Note that all tests and half-tests are placed on the scale of $T_1$. The linkings to the $T_1$ scale are necessary to place all tests and half-tests on a single scale. $T_2$ is linked to $T_1$ through one equivalent-groups link, while $E_2$ and $H_2$ were placed on the $T_1$ scale in two steps, an equivalent-groups link to an exam of equal difficulty ($E_1$ or $H_1$), which in turn is linked with $T_1$ via a single-group design.

Several linkings were performed within this basic schema. First, the direct linking for $T_2$ (to $T_1$) and the two-step linkings for $E_2$ and $H_2$ ($E_2$ to $E_1$ to $T_1$; $H_2$ to $H_1$ to $T_1$) were performed with the Total group data. Then the linkings were performed with Black examinees only and with White examinees only. This set of linkings may be called the *full invariance check,* because it involves all the links needed to place $E_2$, $H_2$, and $T_2$ on a common scale (that of $T_1$), and these linkings were performed for the Total group, as well as for the Black and White examinee subgroups.

In addition, the set of linkings were performed again with a mixture of samples in which the equivalent-groups links (e.g., $H_2$ to $H_1$) used subgroup data while the single-group link (e.g., $H_1$ to $T_1$) used Total group data. This hybrid set of links, called the *hybrid invariance check*, was performed to tease out the effects of differences in test difficulty on subpopulation invariance. If there is a lack of subpopulation invariance, does it lie in the $H_1$ and $E_1$ to $T_1$ links, or in the $H_2$ to $H_1$ and $E_2$ to $E_1$ links? Do the links involving the E half-tests behave differently from those involving the H half-tests?

***Full Invariance Linking Check***

Table 1 contains summary statistics pertaining to the full invariance linking check in which single-step linkings of the total test were done within the Total group and each subgroup. This table summarizes the invariance of the linking process for a SAT-Verbal form that was administered in November 2001. Each column of numbers represents how the Total group and the two subgroups of examinees performed on the test. The table has three horizontal blocks, one for each of three score conversions, namely the conversion obtained when all examinees were used in the (single-step) linking, the conversion obtained when only Black examinees were used in the (single-step) linking, and a conversion obtained when only White examinees were used in the (single-step) linking.

The All, Black, and White vertical blocks contain sample size, mean, standard deviation, skewness, and kurtosis. The Black and White blocks also contain three measures of how much the subgroup conversion differs from the Total group conversion (MD, MAD, and RMSD). MD is the average difference between the subgroup and total group conversions, in which positive and negative numbers can cancel each other out. The mean absolute difference (MAD) is the average of the absolute value of the differences between the subgroup and total group conversions at each score level. RESD is the square root of the expected value of the squared differences between the subgroup and Total group conversions. Like MAD, RESD is always positive and greater than or equal to MD.

Generally, an RESD less than 5 on the SAT scale is considered small enough to ignore because SAT scores are reported on a scale in which the score-reporting unit is 10 points. Differences less than half of a score-reporting unit are considered small enough to ignore because they are too small to be noticed. The only RESD value that comes close to 5 points is the Black group conversion in the Black group. Had the Black group conversion been used in place of the Total group conversion, the average score for Black examinees would have been 432.29 instead of the 435.76 obtained with the Total group conversion. It should be noted that the White group would have been about 2 points lower at the mean had its group-specific conversion been used. As noted before, the difference measures are small and invariance holds for this set of links.

**Table 1**

***Summary Statistics for Full Invariance Check on Total Test Linking***

| Total test | | Sample | | |
|---|---|---|---|---|
| | | All | Black | White |
| | N | 133,120 | 13,659 | 69,698 |
| Total conversion | Mean | 494.64 | 435.76 | 517.12 |
| | Std | 104.67 | 94.06 | 94.31 |
| | Skewness | -0.00 | 0.18 | 0.08 |
| | Kurtosis | 3.02 | 3.07 | 3.12 |
| Black conversion | Mean | 492.71 | 432.29 | 515.81 |
| | Std | 107.13 | 96.73 | 96.33 |
| | Skewness | -0.03 | 0.19 | 0.12 |
| | Kurtosis | 2.96 | 3.14 | 3.19 |
| | MAD | 2.51 | 3.63 | 2.01 |
| | RESD | 3.52 | 4.61 | 2.85 |
| | MD | -1.92 | -3.47 | -1.31 |
| White conversion | Mean | 492.83 | 433.81 | 515.39 |
| | Std | 104.75 | 94.54 | 94.18 |
| | Skewness | -0.03 | 0.19 | 0.11 |
| | Kurtosis | 3.00 | 3.10 | 3.14 |
| | MAD | 1.81 | 1.95 | 1.72 |
| | RESD | 2.03 | 2.18 | 1.91 |
| | MD | -1.81 | -1.95 | -1.72 |

Table 2 presents the results of the full invariance check for the hard half tests. Using Figure 1, this linking is $H_2$ to $H_1$ to $T_1$, which is two-step linking. All examinees were used in establishing the conversion from $H_2$ to $H_1$ to $T_1$. Then this two-step process was repeated with Black examinees only, and then with White examinees only. The format of Table 2 is the same as that of Table 1.

In contrast to the total test, the hard half-test linkings are sensitive to the group used to determine the conversion. Note that all three RESDs for the Black group conversion are greater than 5 in magnitude, and that all three conversions are lower than that of the Total group conversion, as indicated by the negative sign for MD. The results for the White group conversion also differ considerably from the Total group conversion, especially in the Black sample, where using the White group conversion increases the Black group mean by 7 points while reducing the standard deviation by 7 points. Also note that the standard deviations for the White group conversion are lower than for the Total group conversion. This use of the White group conversion for Black examinee scores is similar in a way to how Freedle (2003) used the White group regression with Black examinee scores, but with one major exception: Here, an equating is being done, whereas Freedle used inverse regression.[1]

**Table 2**

***Summary Statistics for Full Invariance Check on Hard Half-Test Linking***

| | | Sample | | |
|---|---|---|---|---|
| Hard half-test | | All | Black | White |
| Total conversion | N | 133,120 | 13,659 | 69,698 |
| | Mean | 494.72 | 440.43 | 512.91 |
| | Std | 104.33 | 93.24 | 98.90 |
| | Skewness | 0.01 | 0.12 | -0.04 |
| | Kurtosis | 3.00 | 3.19 | 3.14 |
| Black conversion | Mean | 488.85 | 432.28 | 507.80 |
| | Std | 108.37 | 96.54 | 102.87 |
| | Skewness | 0.02 | 0.18 | -0.04 |
| | Kurtosis | 2.87 | 3.07 | 3.00 |
| | MAD | 6.16 | 8.22 | 5.46 |
| | RESD | 7.41 | 9.08 | 6.73 |
| | MD | -5.87 | -8.15 | -5.11 |
| White conversion | Mean | 498.49 | 447.46 | 515.48 |
| | Std | 98.42 | 86.03 | 93.98 |
| | Skewness | 0.13 | 0.24 | 0.09 |
| | Kurtosis | 3.02 | 3.42 | 3.10 |
| | MAD | 4.53 | 7.29 | 3.47 |
| | RESD | 8.07 | 10.98 | 6.56 |
| | MD | 3.77 | 7.03 | 2.57 |

On the basis of the results shown in Table 2, it seems that the linking of the hard half-test ($H_2$) to the total test ($T_1$) is group-sensitive.

The results for the full invariance check of the easy half-test linking are contained in Table 3. Using Figure 1, this linking is $E_2$ to $E_1$ to $T_1$, which is two-step linking. All examinees were used in establishing the conversion from $E_2$ to $E_1$ to $T_1$. Then this two-step process was repeated using only Black examinees, and then using only White examinees. The format of Table 3 is the same as that of Table 1.

**Table 3**

***Summary Statistics for Full Invariance Check on Easy Half-Test Linking***

| | | Sample | | |
|---|---|---|---|---|
| Easy half-test | | All | Black | White |
| Total conversion | N | 133,120 | 13,659 | 69,698 |
| | Mean | 494.00 | 437.27 | 517.22 |
| | Std | 102.75 | 94.53 | 91.85 |
| | Skewness | -0.13 | 0.18 | -0.05 |
| | Kurtosis | 2.75 | 2.99 | 2.71 |
| Black conversion | Mean | 489.46 | 432.12 | 513.00 |
| | Std | 103.45 | 96.16 | 91.95 |
| | MAD | 4.54 | 5.15 | 4.22 |
| | RESD | 5.16 | 5.77 | 4.76 |
| | MD | -4.54 | -5.15 | -4.22 |
| | Skewness | -0.19 | 0.14 | -0.12 |
| | Kurtosis | 2.73 | 2.90 | 2.71 |
| White conversion | Mean | 491.24 | 434.12 | 514.69 |
| | Std | 103.29 | 95.82 | 91.85 |
| | MAD | 2.77 | 3.15 | 2.53 |
| | RESD | 3.28 | 3.90 | 2.79 |
| | MD | -2.77 | -3.15 | -2.53 |
| | Skewness | -0.18 | 0.13 | -0.10 |
| | Kurtosis | 2.78 | 2.97 | 2.75 |

As in the previous tables, the RESD is the measure of subpopulation invariance or subpopulation sensitivity. The RESD values for the Black group and the Total group are just above and below the critical value of 5 when the Black group conversion is compared to the

Total group conversion. The RESD values for the White group conversion are less than 4 points for all groups. Linkings for this half-test are less population-sensitive than those for the hard half-test, perhaps because the reliability is higher, as seen in Table 4, or because the score distributions for the easy half-test are less bunched, as seen in Table 5.

Table 4 contains reliability information for one verbal test administered in November 2001 and for the hard and easy half-tests composed of the 39 hardest items and the 39 easiest items from that test, respectively. The reliability is an estimate of how highly correlated two tests (such as two hard half-tests) would be with each other, that is, how similar the rank ordering of the same examinees would be across different editions of the same test. Note how less reliable the hard half-test is for the Black group—.83. This low reliability for Black examinees on the hard half-test reflects the inappropriateness of this test for this group. As expected, the easy half-test is considerably more reliable than the hard half-test for the lower-scoring Black group—.89.

**Table 4**

***Reliabilities in Different Subgroups for an SAT-Verbal Test in November 2001—Easy and Hard Half-Tests***

| | Verbal | | |
|---|---|---|---|
| | Total<br>78 items | Hard<br>39 items | Easy<br>39 items |
| Blacks | .92 | .83 | .89 |
| Whites | .93 | .87 | .88 |
| Other Ethnicity | .94 | .89 | 91 |
| Total Group | .94 | .88 | .90 |
| Females | .94 | .87 | .90 |
| Males | .94 | .88 | .91 |

Table 5 contains five-number summaries, and means, standard deviations, skewness, and kurtosis of score distributions of different groups on the total test, and the hard half-test and easy half-test from an SAT-Verbal form administered in 2001. The first horizontal block of numbers contains numbers for the Total group, followed by horizontal blocks for Black examinees and White examinees. The top number within each block is the number of examinees in the group. Then, below this, come the means, standard deviations, skewness, and kurtosis, followed by the five-number summaries. The first number of a particular five-number summary is the score that divides the group of examinees into the highest 5% and everyone else. The second number splits the group into the highest 25% and everyone else. The middle number is the median, the score that divides the group into the highest scoring and lowest scoring halves. The next number separates the lowest 25% from the rest. The last number in the set is the score that divides the group into the lowest 5% and everyone else.

**Table 5**

***Raw Score Distributions for SAT-Verbal Total Test and Half-Tests***

| | | Total 78 items | Hard 39 items | Easy 39 items |
|---|---|---|---|---|
| Total examinees | N | 133,120 | 133,120 | 133,120 |
| | Mean | 35.5 | 10 | 25.6 |
| | Std | 16.9 | 9.2 | 8.9 |
| | Skewness | 0.09 | 0.73 | -0.69 |
| | Kurtosis | -0.64 | -0.14 | -0.12 |
| | 95% | 64 | 28 | 38 |
| | 75% | 48 | 16 | 33 |
| | Median | 35 | 8 | 27 |
| | 25% | 23 | 3 | 20 |
| | 5% | 8 | -2 | 8 |
| Black examinees | N | 13,659 | 13,659 | 13,659 |
| | Mean | 25.8 | 5.4 | 20.5 |
| | Std | 14.8 | 7 | 9 |
| | Skewness | 0.52 | 1.26 | -0.16 |
| | Kurtosis | -0.11 | 1.73 | -0.67 |
| | 95% | 53 | 20 | 34 |
| | 75% | 35 | 9 | 27 |
| | Median | 24 | 4 | 21 |
| | 25% | 15 | 1 | 14 |
| | 5% | 4 | -3 | 5 |
| White examinees | N | 69,698 | 69,698 | 69,698 |
| | Mean | 39.1 | 11.5 | 27.7 |
| | Std | 15.5 | 9.1 | 7.5 |
| | Skewness | 0.06 | 0.60 | -0.76 |
| | Kurtosis | -0.57 | -0.32 | 0.24 |
| | 95% | 66 | 29 | 38 |
| | 75% | 50 | 17 | 33 |
| | Median | 39 | 10 | 29 |
| | 25% | 28 | 4 | 23 |
| | 5% | 14 | -1 | 13 |

The most striking feature of Table 5 is the stark contrast between the easy half-test and the hard half-test. For the Total group, the median score is 8 on the hard half-test, and the middle half of the scores fall between 3 and 16. In contrast, the median is 27 on the easy half-test, and the middle half of the scores fall between 20 and 33. For the Black group, the middle half of the scores on the hard half-test fall between 1 and 9, with a median of 4, while the median on the easy half-test is 21, with midrange of 14 and 27. Nearly 25% of the Black examinees score below 0 on the hard half-test, which is the score that would be expected for an examinee who guessed randomly on every question. Clearly, these half-tests have very different statistical properties from each other, and from the total test.

The hard half-test seems to work in the higher proficiency region. By this we mean that the number of raw score points is large in comparison to the proportion of the examinees who fall within this region, which implies that a change in raw scores in that region is making relatively fine distinctions among examinees. Note that in the Total group for the hard half-test, the top 25% of the examinees scored at or above 16 and the top 5% scored at or above 28, a difference of 12 raw score points for 20% of the examinees. For Black examinees, these numbers are 9 and 20, a difference of 11 raw score points for 20% of the examinees. At the low end, however, the hard half-test is much less effective because small discriminations in raw score points are being made with large groups of examinees. In the Total group, the difference between the 5% and 25% is 5 raw score points [3–(-2)], and for the Black group, it is 4 raw score points [1–(-3)]. The Freedle hard half-test actually fails to measure effectively in the proficiency region where Freedle claims it works best at elevating scores.

In contrast, the easy half-test uses 12 (20–8) raw score points between the 5% and 25% percentile in the Total group, and 9 (14–5) raw score points in the Black group. At the high proficiency regions, the easy half-test is much less effective, spending only 5 (38–33) raw score points to measure in the 75% to 95% region in the Total group, and 7 (34–27) raw score points in the Black group.

The bunching of scores at the low end on the hard half-test makes it difficult to match these score distributions to the more spread out score distributions for the total test in a way that is consistent across subpopulations. In addition, the mismatch of difficulty with population proficiency makes matters worse because the hard half-test is less reliable, which means it ranks examinees less consistently across different editions of the same test. The low reliability of the hard half-test in the Black subpopulation attests to this and to the psychometric unsoundness of Freedle's hard half-test idea (Dorans & Zeller, 2004).

***Hybrid Linking Check***

What is really causing this group sensitivity? Is it caused by the fact that the hard half-test linkings are group dependent? Or is it a result of the within-test single-group linkings? Recall from Figure 1 that the hard half-test linking is a composite of two separate linkings — an equivalent-groups linking of hard halves, and a single-group linking of a hard half-test to a middle-difficulty total test. What happens if we replace the single-group-link subgroup conversion with the total-group conversion? This use of a common conversion for the single-group hard half-test to total test link would enable us to check if the hard half-test to hard half-test equivalent-group link is invariant.

The score distribution summaries in Table 5 show that the hard half-test has a very strange distribution: 75% of the scores in the Total group are at or below a score of 16 (17 score points [0 to 16] is 42.5% of 40 score points, the number of nonnegative integer scores that can be obtained on a 39-item hard half-test), and 25% are at or below 3 (10% of the number of nonnegative integer scores). In contrast, on the total test, 75% of the Total group scores are at or below 48 (61% of the number of nonnegative integer scores on a 78-item total) and 25% are at or below 23 (29% of the number of nonnegative integer scores).

Table 6 contains the results of the hybrid linking of a hard half-test to a total test through another hard half-test. Using Figure 1, this linking is $H_2$ to $H_1$ to $T_1$. Data from the group (all examinees) or subgroup (Black-only or White-only examinees) of interest are used to establish an equivalent-groups link from $H_2$ to $H_1$. The All examinees group is used every time to establish the single-group link of $H_1$ to $T_1$. This hybrid linkage permits us to tease out the effects of test difficulty from those of subgroup invariance.

When compared to Table 2, the results in Table 6 suggest that much of the population sensitivity to subgroup may in fact be attributed to the single-group linking of the hard half-test to the total test. This is true for White examinees and, to a lesser extent, for Black examinees. All REMSD values are less than 5. The larger standard deviation for the Black group conversion and the smaller standard deviation for the White group conversion suggest that there is still some population sensitivity that could affect scores in the tails of the distributions.

**Table 6**

***Summary Statistics for Hybrid Invariance Check on Hard Half-Test Linking***

| | | Sample | | |
|---|---|---|---|---|
| Hard half-test | | All | Black | White |
| | N | 133,120 | 13,659 | 69,698 |
| Total conversion | Mean | 494.72 | 440.43 | 512.91 |
| | Std | 104.33 | 93.24 | 98.90 |
| | Skewness | 0.01 | 0.12 | -0.04 |
| | Kurtosis | 3.00 | 3.19 | 3.14 |
| Black conversion | Mean | 494.66 | 438.19 | 513.62 |
| | Std | 108.13 | 97.25 | 102.32 |
| | MAD | 3.64 | 4.02 | 3.47 |
| | RESD | 4.39 | 4.88 | 4.14 |
| | MD | -0.06 | -2.25 | 0.72 |
| | Skewness | -0.04 | 0.12 | -0.10 |
| | Kurtosis | 2.89 | 3.07 | 3.04 |
| White conversion | Mean | 493.54 | 440.00 | 511.47 |
| | Std | 102.93 | 91.91 | 97.61 |
| | MAD | 1.61 | 1.23 | 1.72 |
| | RESD | 1.92 | 1.56 | 2.01 |
| | MD | -1.18 | -0.43 | -1.44 |
| | Skewness | 0.01 | 0.12 | -0.03 |
| | Kurtosis | 3.01 | 3.23 | 3.15 |

## Summary of Score Linking Checks

Freedle made claims about the fairness of SAT scores. His empirical justification was based on flawed scoring and inappropriate DIF analyses (Dorans & Zeller, 2004). His solution involved inappropriate linking of scores on a hard half-test to scores on the SAT scale. Score equity assessment (SEA) is a direct way of evaluating the comparability of scores on a hard half-test to scores based on a full-length test that contains a full range of difficulty.

The SEA analysis used in this study suggests that scores on a hard half-test may be equatable across hard half-tests, but that the hard half-test to total test linking may be population-dependent. For this study, it is fair to say that the scores produced on the hard half-test cannot be used interchangeably with scores produced on the full-length SAT-Verbal test. As a consequence, the meaning assigned to full-length SAT-Verbal test scores cannot be transferred to scores from the hard half-test.

There are at least two reasons for this lack of transferability. First, less-proficient examinees would prefer taking the harder test that produces less-reliable scores, because it is more of a lottery than an evaluation of their capabilities. Getting lucky can help the less proficient more than the proficient. Second, the peculiar nature of hard half-test raw scores cannot be made to look like scores on the total test. Distributions of scores based on tests with very different difficulty levels are too divergent to be brought together.

## References


von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004). *The kernel method of test equating*. New York: Springer.

Dorans, N. J. (2004a). Freedle's table 2: Fact or fiction. *Harvard Educational Review, 74*(*1).*

Dorans, N. J. (2004b). Using population invariance to assess test score equity. *Journal of Educational Measurement, 41*(1), 43-68.

Dorans, N. J., & Zeller, K. (2004). *Examining Freedle's claims about bias and his proposed solution: Dated data, inappropriate measurement, and incorrect and unfair scoring* (ETS RR-04-26). Princeton, NJ: ETS.

Freedle, R. O. (2003). Correcting the SAT's ethnic and social bias: A method for reestimating SAT scores. *Harvard Educational Review, 73,* 1-43.

Holland, P. W., & Wainer, H. (Eds.) (1993). *Differential item functioning*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York: Springer-Verlag.

Mathews, J. (2003, November). The bias question. *The Atlantic Monthly.* 130-140.

Schmitt, A. P., Holland, P. W., & Dorans, N. J. (1993). Evaluating of hypotheses about differential item functioning. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 281-315). Hillsdale, NJ: Erlbaum Associates.

## Notes

[1] Regression is a statistical term for the procedure for finding the average or expected score on one variable (in this case, the Hard Test score) for each level of a second variable (in this case, the Total Test score). Typically, such a regression could be used to predict the value of the second variable, given the value of the first variable. The process of using such a regression backwards, to predict the first variable given the value of the second variable, is known as inverse regression.