Research Report

Examining Freedle's Claims About Bias and His Proposed Solution: Dated Data, Inappropriate Measurement, and Incorrect and Unfair Scoring

Neil J. Dorans

Karin Zeller

Research & Development

July 2004
RR-04-26

Examining Freedle's Claims About Bias and His Proposed Solution: Dated Data, Inappropriate Measurement, and Incorrect and Unfair Scoring

Neil J. Dorans and Karin Zeller

ETS, Princeton, NJ

July 2004

Abstract

In the Spring 2003 issue of *Harvard Educational Review*, Roy Freedle stated that the SAT® is both culturally and statistically biased, and he proposed a solution to ameliorate this bias. His claims, which garnered national attention, were based on serious errors in his analysis. We begin our analyses by assessing the psychometric properties of Freedle's recommended hard half-test that he thinks should form the basis for the supplemental SAT score he proposes to report. Next we demonstrate his justification for a score based on this half-test is based on a flawed analysis. The numbers in his critical Table 2 do not represent what he claims they do. We then demonstrate what occurs when current data are used with both the correct scoring and the incorrect scoring proposed by Freedle. When the table is constructed correctly using current data, the effects that Freedle reported are reduced substantially in magnitude to the point where they do not warrant any of the corrective actions he proposes. We conclude our analysis of Freedle's claims with DIF analyses of our own that are inconsistent with his DIF analyses of old SAT test editions he used as his initial motivation for correcting the SAT scoring procedure.

Key words: Fairness, differential item functioning; computing averages, psychometric properties of tests

Acknowledgements

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

1. Overview

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 then goes on to propose a solution to the problem by, as the title of his article indicates, "Correcting the SAT's Ethnic and Social-class Bias: A Method for Reestimating SAT Scores."

Freedle's claims have garnered national attention via an article in the *Atlantic Monthly* (Mathews, 2003). His claims, however, are based on serious errors in his calculations (Dorans, 2004). He incorrectly used results from his variant of a statistical procedure (Dorans & Kulick, 1986) developed 20 years ago for assessing the statistical fairness of test questions. What did he do wrong? How did he misuse the standardization procedure for doing Differential Item Functioning (DIF)? What happens when his scoring error is corrected?

Freedle proposed a solution to the alleged statistical and bias he claimed to have found. As just mentioned, his claims were based on serious errors in his analyses. We begin our analyses in Section 2 by assessing the psychometric properties of Freedle's recommended hard half-test that formed the basis for the supplement score that he proposes to report. In Section 3, we demonstrate that his justification for a score based on this half test is based on a flawed analysis because the numbers in his critical Table 2 do not represent what he claims they do. We then demonstrate what occurs when current data is used with both the correct scoring and the incorrect scoring proposed by Freedle. When the table is constructed correctly using current data, the effects that Freedle reported are reduced substantially in magnitude to the point where they do not warrant any of the corrective actions he proposes. We conclude our analysis in Section 4 with DIF analyses of our own that are inconsistent with his DIF analyses of old SAT test editions, which he used as his initial motivation for correcting the SAT scoring procedure. Section 5 summarizes our findings and mentions potential avenues for future research.

2. The Freedle Hard Half-test Suggestion

2.1 Difficulty Considerations

Freedle recommended supplementing regular SAT scores with scores based on a half test composed of the hardest SAT questions on a particular form (called R-scores by Freedle), because he states it would give higher and more appropriate scores for non-White examinees. Specifically, he looked at Black and White examinees in his analysis.

A basic tenet of sound measurement is to target the difficulty of the test to the level of the examinees who comprise the test-taking population. This principle pervades all measurement. The tool used to measure should depend on what it is you seek to measure. Suppose you wanted to measure the distance between two city blocks that were quite a distance apart. You might be inclined to hop in your car and use the odometer, which measures in tenths of miles. Now, suppose you wanted to measure the length of your shoe. A ruler would be appropriate. The measuring device in a shoe store would work even better if you wanted to know the size of the shoe. Would you place a shoe alongside your car and use the odometer to measure it its length? Of course not.

Freedle's unusual recommendation runs counter to accepted wisdom that tests targeted to the ability level of the examinees being tested will tend to produce the most appropriate scores for that population. This principle is what motivated many of the advocates of computerized adaptive testing. Although it has been difficult to sustain continuous adaptive testing in practice, the fundamental idea behind adaptive testing is simple and sound: Do not waste examinee time with questions that are too hard or too easy for them because these questions do not provide us with useful information. Just as you would want to match the measuring device to the distance you want to measure, educational measurement professionals would like to match the test's questions to the level of the examinees if possible. Computer adaptive testing gives harder questions to examinees who correctly answer items and easier questions to examinees who answer items incorrectly. Freedle turns this principle upside down by, in essence, urging that harder questions be given to groups that have a larger proportion of examinees who tend to answer more questions incorrectly because they will get higher scores that way. Freedle suggests that Black examinees be given scores based on hard tests. What properties would these scores have?

Table 1 contains five-number summaries; means, standard deviations, skewness, and kurtosis of score distributions of different groups on the total test; and the hard half- and easy

half-tests from an SAT Verbal form administered in 2001. The first horizontal block of numbers contains numbers for the total group, followed by Blacks and Whites. The top number in each column 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 five-number summaries 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.

The most striking feature of Table 1 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 to 27. Nearly 25% of the Black examinees score below zero, which is the score you would be expected to get if you 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 to the proportion of the test takers who fall within this region is large, which implies that a change in raw scores in that region is making relatively fine distinctions among examinees. Note that in the total group, the top 25% of the students score at or above 16 and the top 5% score at or above 28 on the hard half-test, a difference of 12 raw score points for 20% of the students. For Black examinees, these numbers are 9 and 20, a difference of 11 raw score points for 20% of the people. 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 fails to measure effectively where Freedle claims it works best at elevating scores.

Table 1

Raw Score Distributions for Verbal Total Test and Half-tests

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

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 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.

In sum, the hard half-test is more effective than the easy half-test at distinguishing among students with high levels of verbal reasoning proficiency, while the easy half-test does a better job at the lower end. The total test, which is the sum of the easy and hard half-tests does better than either alone. Except for the highest levels of verbal reasoning proficiency, the hard half-test or Freedle test is ill-suited for lower-scoring members of the Black group. It is especially inappropriate at the score levels where Freedle made his strongest claims.

2.2 Reliability Considerations

Table 2 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, respectively. The reliability is an estimate of how highly correlated two hard half-tests would be with each other, that is, how similar rank ordering of the same people would be across the different editions of the same test. Note how less reliable the hard half-tests are for the Black group, 0.83. Reliabilities in the high 0.80s or the 0.90s are desirable for tests that produce scores for individual examines that are used in high-stakes settings. The low reliabilities for Black examinees on the hard half-test reflect the inappropriateness of these tests for this group. Freedle's suggestion is bad practice. As expected the easy tests are considerably more reliable than the hard half-tests in the lower-scoring Black group.

In sum, there is little empirical justification for Freedle's suggestion. As will be seen in Section 4, the DIF analyses do not support it. In fact, there are strong arguments against it. Hard half-tests are poor measures for lower-scoring examinees. While Freedle's hard half-tests might work well among higher scoring students, they will perform poorly for lower-scoring students, including many Black examinees. Nearly 25% of Black examinees receive negative below-chance scores on the hard half-test.

Table 2

Reliabilities in Different Subgroups for a SAT Verbal Test in November 2001—Its Easy and Hard Half-tests

		Verbal	
	Total 78 items	Hard 39 items	Easy 39 items
Blacks	0.92	0.83	0.89
Whites	0.93	0.87	0.88
Other ethnic	0.94	0.89	0.91
Total group	0.94	0.88	0.90
Females	0.94	0.87	0.90
Males	0.94	0.88	0.91

3. Freedle's Flawed Averaging

Freedle's Table 2, which is based on selected data from a 1980 SAT test edition, is central to his argument. Based on the numbers in this table, he argues the Black examinees perform better on the hard questions than do White examinees that are matched to the Black examinees on total test performance. He then argues that if scores of Black examinees were based on the hard questions only, subgroup differences would be reduced considerably.

The correct way of combining question data across a collection of items is to define a group of examinees, see how that group performed on each question, and sum that performance across questions. In Freedle's case, he was interested in two groups, a Black group and a White group. Both groups were administered questions on a 1980 SAT Verbal form.[1] It appears that Freedle tried constructing a table that describes how well members of the Black group and members of the White group who have the same scores on the total test could be expected to do on a subtest composed of the hardest questions on the total test. He failed to accomplish that goal. Instead, Freedle created a table of numbers that does not describe how any group performed on a subtest composed of the hardest items on the total test. Freedle found differences in this ill-constructed table and posited explanations for why these differences occurred.

Freedle used neither the same Black group across questions nor the same White group across questions to construct his table. Instead, the groups that went into the construction of the table varied from question to question. Examinees who did not respond to a question for one reason or another were excluded from the performance data for that question.

3.1 Freedle's Incorrect Table 2

As mentioned, the core of Freedle's argument resides in his Table 2, which has been inserted here as Table 3. The first column of Table 3 contains scaled scores on the full SAT Verbal form. These scaled scores are the result of a scoring and equating process that adheres to the basic requirements of equating described in Kolen and Brennan (1995). A primary goal of equating is to produce scores that enable examinees to be compared fairly across different editions of a test.

Table 3

Reproduction of Freedle's (2003) TABLE 2

Percentage of correct responses to the 40 hardest verbal items					
Original SAT score for 85 verbal items	White examinees % correct (40 hardest)	Black examinees % correct (40 hardest)	Gain score for Black examinees	Average reestimated R-SAT score for Black examinees	Bl-Wh % diff.
200	11.9%	13.0%	10 pts	200 + 10 = 210	1.1%
210	12.9%	14.6%	40 pts	210 + 40 = 250	1.7%
220	13.2%	15.6%	60 pts	220 + 40 = 260	2.4%
230	14.2%	16.1%	60 pts	230 + 60 = 290	1.9%
240	14.5%	16.5%	50 pts	240 + 50 = 290	2.0%
250	14.6%	17.2%	50 pts	250 + 50 = 300	2.6%
260	15.1%	17.6%	40 pts	260 + 40 = 300	2.5%
270	15.2%	18.2%	40 pts	270 + 40 = 310	3.0%
280	15.9%	18.4%	30 pts	280 + 30 = 310	2.5%
290	16.1%	19.0%	40 pts	290 + 40 = 330	2.9%
300	17.4%	19.2%	30 pts	300 + 30 = 330	1.8%
310	18.3%	20.5%	40 pts	310 + 40 = 350	2.2%
320	18.1%	21.1%	40 pts	320 + 40 = 360	3.0%
…					
470	33.9%	35.3%	10 pts	470 + 10 = 480	1.4%
480	35.6%	36.4%	10 pts	480 + 10 = 490	0.8%
490	36.5%	37.7%	10 pts	490 + 10 = 500	1.2%
…					
640	68.0%	69.2%	10 pts	640 + 10 = 650	1.2%
…					
800	100%	100%	0 pts	800 + 00 = 800	0.0%

The second column is the "percentage correct" for White examinees on the 40 hardest items, while the next column is the "percentage correct" for Black examinees on these 40 items.

The "percentage correct" is in quotes intentionally. Freedle obtained these values indirectly, as noted in his footnotes. The 4th column in Table 3 is a so-called score gain that is obtained by finding the closest predicted values on the hard half-test in columns 2 and 3, and using inverse regression[2] to obtain a score that appears to be on the 200–800 scale. For example, Freedle considers 16.1% in column 2 and 16.1% in column 3 to be equivalent. He goes into the regression backwards and obtains scores of 230 for Black examinees and 290 for White examinees. He takes the difference between these values, 60, which is in the 4th column. He then adds the 60 to scaled score of 230 for Black examinees to get a new score of 290 in column 5, which Freedle calls his R-score. Finally, the 6th column is simply the difference in percentage correct on the half-test between Black examinees and White examinees.

In Section 3.3, we address problems associated with the 4th through 6th columns of Table 3, but for now we want to focus on columns 2 and 3. The following demonstrates what Freedle did with these columns:

Let N_i represent the number of people or examinees that were administered question i. Everyone was administered the same set of questions, so N_i is a constant across all items. Let N denote this constant, the number of examinees administered the test. For each question i, the number of examinees N is the sum of four groups of examinees: R_i (the group whose answers were Right); W_i (the group whose answers were Wrong); O_i (the group that appears to have chosen to Omit the question); and NR_i (the group that appears to have Not Reached the item and therefore would not have had the opportunity to answer or to omit the question). For every question, $N = R_i+W_i+O_i+NR_i$. Across questions, the values of the component parts change. For hard questions, R_i is small while W_i and O_i tend to be large; for easy questions R_i is large. NR_i tends to be 0 or small until the end of a timed section.

For question i, Freedle used $R_i /(R_i + W_i)$ to define "percentage correct." Instead of using the same group of examinees for each question, Freedle used data only from individuals who answered the item correctly or incorrectly, and thereby excluded those examinees who did not respond to the question. On a test like the SAT, it is often wise not to respond to very hard questions because nonresponse nets a higher score than an incorrect response. Many people opt to skip hard questions. In contrast, nearly everyone answers the easy questions.

For example, on an SAT I Verbal test edition administered in November 2001, 12,522 Black examinees supplied data for DIF analysis on a verbal test composed of 78 questions. On

the first question in the first section, 99% of the Black examinees answered the question. By the 5th question, only 90% answered the question, and 10% did not. By the 8th question, only 75% answered the question. And only 52% of the Black examinees answered the 22nd question in the section. So on item 1, nearly all 12,522 Black examinees contribute to $R_1/(R_1 + W_1)$. On item 22, only slightly more than half contribute to $R_{22}/(R_{22} + W_{22})$.

Note that the sum $(R_i + W_i)$ varies from question to question. As a consequence, the numbers do not describe how any particular group, be it Blacks or Whites, performed on the hard half-test. For example, what does the sum of $R_1/(R_1 + W_1)$ and $R_{22}/(R_{22} + W_{22})$ mean for the Black group? This sum does not represent how the total Black group would perform on a mini-test composed of questions 1 and 22. Additionally, it does not describe the performance of a group defined by the sum of $R_1 + W_1$ (i.e., of those who answered Question 1). It also does not describe the performance of a group defined by the sum of $R_{22} + W_{22}$ (i.e., those who answered Question 22). We cannot define what this sum describes in easily understood terms. Imagine trying to define that sum when you add in the remaining 76 SAT Verbal questions, keeping in mind that the individuals who omit one question are not always the same as, or a subset of, the individuals who omit other questions.

Why did Freedle use $R_i/(R_i + W_i)$ to define "percentage correct"? Freedle considered three variants of the standardization procedure for DIF analysis (Dorans & Kulick, 1986), as he discussed in his Appendix A. The major problem with using the particular variant he chose, which makes use of $R_i/(R_i + W_i)$, is that the group used to study a question varies from question to question in a systematic way related to the difficulty of the question and the proficiency of the examinees taking the question. Summing these DIF values across questions produces an "average" that is not associated with the performance of any single group or subgroups of the total group.

A preferred alternative for DIF is described in Dorans and Holland (1993): "Score the items in a manner that is consistent with the total score. Use item formula scoring and use everyone in both groups." This is equivalent to an even more direct approach: "Formula score the hard test directly, and determine its relationship to the total test score across all examinees. When this correct set of calculations is performed, a correct table is created and the effects of the dubious R-score almost vanish."

The SAT is a formula-scored test. The examinees took the test under formula scored conditions, and the total score was based on formula scoring of the questions. The hard half-test should have been formula-scored using data from all examinees, including those who did not respond. Typically, formula scoring employs a rule (fs^-) that assigns plus 1 to a correct response, a 0 to a nonresponse (omit or not reached), and minus a small fraction[3] to an incorrect response such that the maximum score is 1 and the minimum is minus the small fraction. Here we use an alternative but equivalent formula scoring rule (fs^+) in which the maximum score is 1 but the minimum is 0. To make the number look like a percentage, we multiple it by 100. The numerator is an item formula score in which a 1 is assigned to the R_i, a small fraction, $1/k$, where k is the number of options, is assigned to both the NR_i and O_i, groups, and a 0 is assigned to the W_i; these numbers are summed and divided by $(R_i+W_i+O_i+NR_i)$. With this formula, the denominator is all examinees $(R_i+W_i+O_i+NR_i)$ at the score level under consideration, so the same group of examinees is used to calculate the DIF measure. This measure of item performance uses all the data, and treats the questions in a manner that is mathematically consistent with how the total test is scored because $fs^+ = (fs^-)*(k-1)/k + 1/k$.

A simple illustration can be used to demonstrate the relationships among different types of scores. In Table 4, we have five people and five questions. Person A answers all five questions correctly, while Person E does not answer any correctly.

Table 4

The Responses of Five People on a 5-question Exam: Right, Wrong, Omit, Not Reached

	Item 1	Item 2	Item 3	Item 4	Item 5
Person A	R	R	R	R	R
Person B	R	R	R	R	NR
Person C	R	O	O	W	W
Person D	R	O	W	NR	NR
Person E	W	O	O	W	NR

In Table 5, we score the questions using the fs^+ rule. Then we compute the average score on each question in percentage of maximum score units by averaging the item scores across each of the five examinees. This number appears in the row labeled %MAX. Question 1 was answered correctly by 4 of 5 (80%), while Question 5 was the hardest with a 32% score.

Table 5

fs$^+$ Scores for Five People on Five Five-choice Questions

	Item 1	Item 2	Item 3	Item 4	Item 5
rs	80%	40%	40%	40%	20%
fs$^+$	80%	52%	48%	44%	32%
frsc	80%	100%	67%	50%	50%

Next, in Table 6, we compare average scores based on the fs$^+$ method, the rights scoring method (rs, in which correct responses are assigned a one and all other responses receive a zero), and the Freedle scoring procedure (called frsc). Unlike the rs and fs$^+$ rules, frsc uses only the examinees who attempted a question (i.e., with responses *R* or *W*) in its estimate of the question's average score or difficulty. We see in Table 6, for example, that only two people gave an answer to Question 2 and both answers were correct. That is why the frsc score for Question 2 is 100%.

Table 6

rs, fs$^+$, frsc Scores for Five People on Five Five-choice Questions

	Item 1	Item 2	Item 3	Item 4	Item 5
Person A	1	1.0	1.0	1.0	1.0
Person B	1	1.0	1.0	1.0	0.2
Person C	1	0.2	0.2	0	0
Person D	1	0.2	0	0.2	0.2
Person E	0	0.2	0.2	0	0.2
%MAX	80%	52%	48%	44%	32%

The three scoring procedures produce quite different results. In Table 6, the three scores only agree on Question 1 because everyone gives a response to this easy question. Note that fscr is the often the most unusual score of the three. On Question 2, rs = 40% of maximum, fs$^+$ = 52% and frsc = 100%; while on Question 3, rs = 40%, fs$^+$= 48%, and frsc = 67%.

The %Max for frsc represents the following: On item 1, it is the average of all five people; on item 2, the average of persons A and B; on item 3, the average of persons A, B, and D; on item 4, the average of persons A, B, C, and E; and on item 5, the average of persons A and C. It represents the average score of all five examinees on only one item.

The assumption might be made that the sum of these frsc averages across the five items is the average of all five people. However, this is false. It is a biased estimate of the difficulty of this five-item test for these five people, and the bias is caused by ignoring the nonresponses instead of scoring them as is done with FS+. As will be seen shortly, this bias in the calculation of these averages is a likely explanation for the unusual results obtained by Freedle.

3.2 A Visual Explanation

How do Freedle R-scores differ from rights scores and formula scores? They are a number right score that excludes any examinees that do not respond to the question. Therefore, they are larger than regular number right scores that treat nonresponse as incorrect. Figures 1 and 2 describe the effects of Freedle scoring in a very precise manner. Figure 1 shows the difference in percents, Black group minus the White group, for each of the four possible outcomes (right, wrong, omit, and not reach) conditioned on attained fs^{-}, at each level of fs^{-}. Figure 1 is based on analysis of an SAT Verbal exam administered in November 2001. We have truncated the chance and below chance scores (0 or less) from the low end. At the top end, scores for which there were fewer than 50 (out of 12,522), Black examinees are truncated because their results might be unstable. As a consequence, the formula scores range from 1 to 59 out of a maximum of 78. Within this range, the results should not be influenced much by chance performance or by small numbers of examinees. Figure 2 compares both the differences in frsc and difference in the number right score with each other using the same x-axis as in Figure 1. Note that the range of the y-axis in Figure 2 is 14 times the size of the range in Figure 1.

There is a positive difference in Figure 1 for omits and not-reached, and a negative difference for percentage right, meaning that the Black group tended to omit and not reach more often than the White group at the same fs^{-} level. On the other hand, there is a negative difference for percentage wrong, meaning Black examinees who responded tended to be Wrong at a lower rate than White examinees who responded. (The Black examinees tended to be right at a lower rate also, but these differences were very minor.) The difference in nonresponse rates means that at each score level, a higher proportion of Black examinees were being excluded from the calculation of the Freedle R-score. And the difference in wrong rates means that the Freedle R-score for Black examinees at a given formula score (or scale score) would be higher than for White examinees. These differences are what Freedle obtained. When you remove the students

who do not respond to an item you end up with a difference in Freedle R-scores (frsc), which is inconsistent with both the formula score and the percentage right score.

These points become clearer when you examine Figure 2 closely. You see varying differences in the frsc between the groups and very small and nearly constant magnitude for differences in percentage right. So, at each fs^- the frsc difference fluctuates based on the group's fs score, and rs does not. These factors are what caused his differences: At each formula score point, the Freedle R-score in the Black group (and White group) used different groups of people for each item on the test. In contrast the rights score and formula score used the same groups across all items. Figure 2 demonstrates what Freedle's form of selective sampling can do to a number right score.

It appears as if Freedle thinks that all examinees at a given fs^- obtained their score in the same way. They don't. Some answer all questions. Some never finish the test. Some omit the hard questions. Table 7 illustrates how variability of response patterns for a common scaled score can result in different scores under different scoring rules fs^+, fs^-, rs, and frsc. Listed are four patterns of rights, wrongs and omitted responses that produce the same fs^-, and fs^+, scores, of 49% and 59%, respectively. Note the variation in frsc and the negative relationship it has with rs. When Freedle started to exclude the portion of examinees who do not respond to a question, he changed the definition of the comparable from all examinees with a given fs^- score to the set of examinees with a given fs^- who either got the question correct or got the question wrong. This selection operation explains some of his findings.

Table 7

An Illustration of Variation in rs and frsc for Fixed Levels of fs^- and fs^+

Counts on a 39-item hard test						
Rights	Wrongs	Omits	fs^-	fs^+	rs	frsc
23	16	0	49%	59%	59%	59%
22	12	5	49%	59%	56%	61%
21	8	10	49%	59%	54%	72%
20	4	15	49%	59%	51%	84%
19	0	20	49%	59%	49%	100%

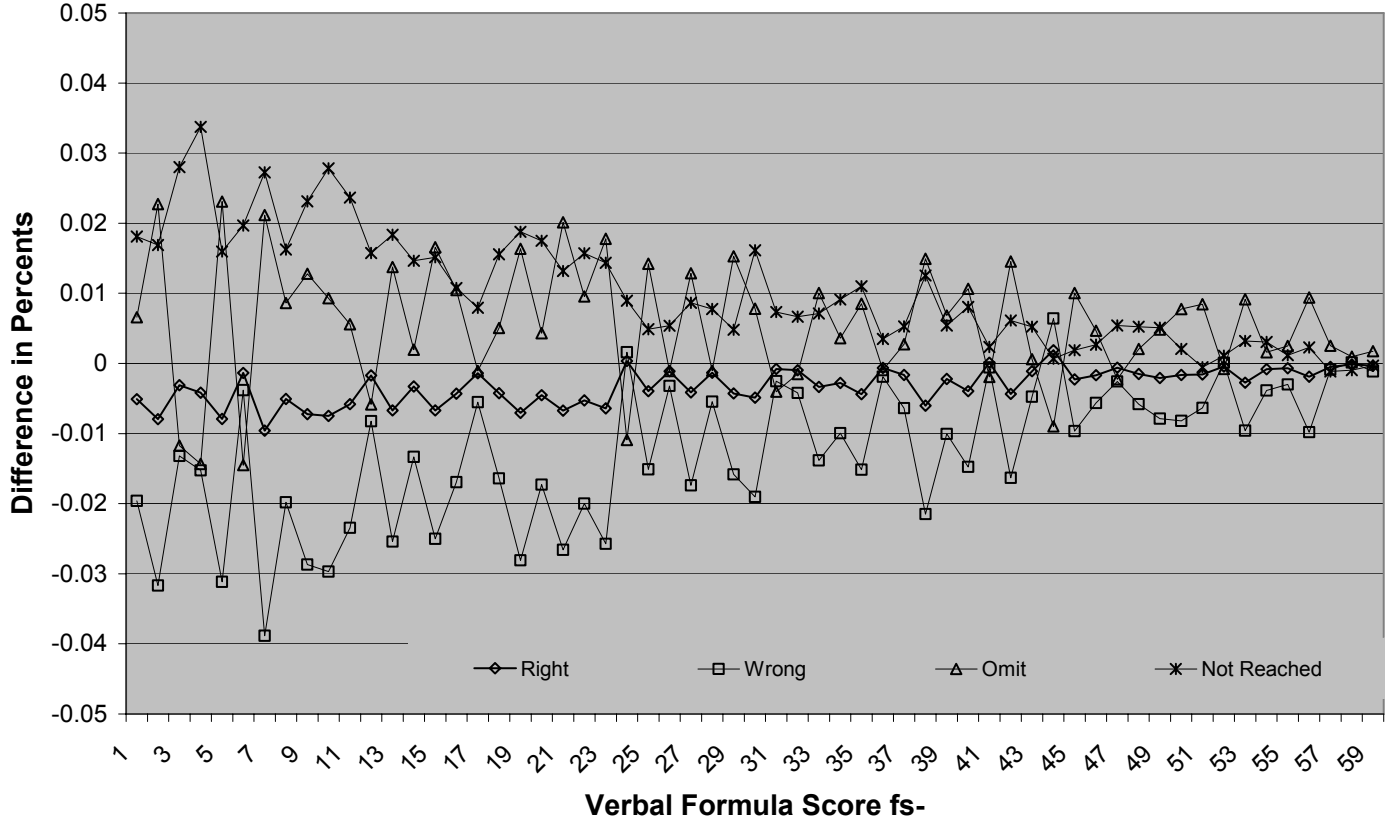

Figure 1. **Differences (Black-White) in percentages by item outcomes (R, W, O, NR).**

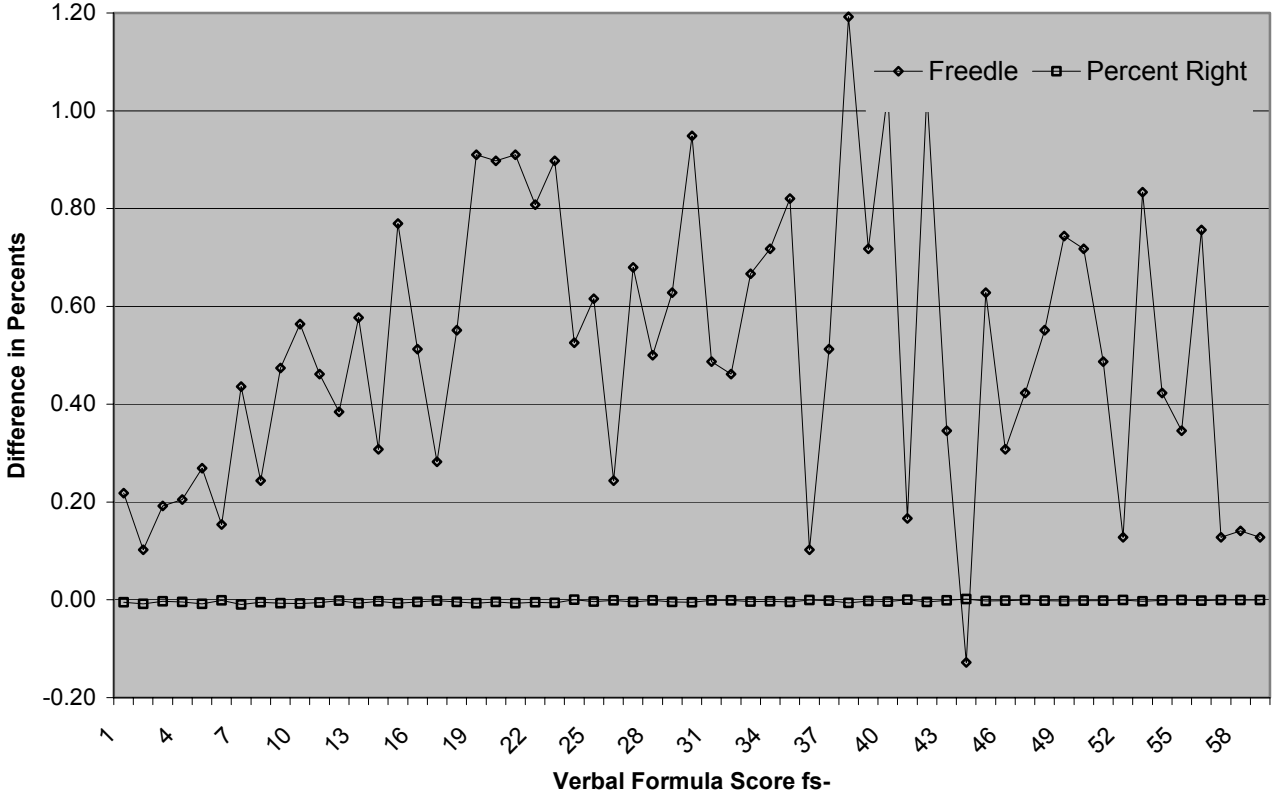

Figure 2. **Differences in percentages: Freedle R-score and rights score.**

3.3 Updated Table 2

What does the hard test/total test relationship look like with a modern SAT I form? What does it look like when proper scoring and improper scoring are compared? Table 8 is based on analysis of an SAT Verbal exam administered in November 2001. A hard half-test composed of the hardest 39 items on the test was evaluated against the full 78-item test to produce this table. The first column is the traditional formula score, (fs^{-}), which has 0 as its chance score on the total test. The range of formula scores is from 1 to 59 as was the case in Figure 1.

The second column in Table 8 contains rounded SAT score in two-digit form (because the third digit is always 0, we decided to place it in parentheses).[4] The score contained in this column differs from Freedle's in two significant ways. First, the scores are on the 1990 SAT I scale (divided by 10), while Freedle's scores were on the old 1941 scale (Dorans, 2002). Second, and more importantly, this table does not throw important information away. Note that many rounded scaled scores are associated with two fs^{-} scores. For example, each rounded scale score from 41 to 50 is associated with two fs^{-} scores. Freedle's analysis lacked this precision and some of his effects may have been exaggerated because of the rounding effect evident in his data.

The next two columns contain the Freedle $R/(R+W)$ proportions correct summed across the 39 hard items on a 0 to 1 scale expressed as percentages. Keep in mind that the groups, Black and White, at each fs^{-} score level that went into this calculation included only those who responded to that question. Note that examines who do not respond to the question do not get a frsc for that question, but since they answer at least one question they get a frsc based on the questions answered.

Although it makes it a little harder to relate subsequent columns to these two columns, they are ordered as Freedle ordered his columns. The Whites column precedes the Blacks column. The next columns are the differences, Black% – White%, which entails subtracting column 3 from column 4. These are an attempt to duplicate what Freedle appears to have done. The next column will be discussed later.

After presenting four columns related to the incorrect scoring procedure, we present four columns for the correct scoring procedure in which all examinees' scores on all 39 hard questions are included. These results are discussed in the next section.

Three aspects of this table are worth noting. First, the properly pooled data reveal that Blacks tend to do slightly better than matched Whites on the hard half of the test. Possible explanations for this are discussed in the summary section.

Second, these differences are much smaller than those reported by Freedle (see Table 3 above). For the correct scoring of current data all but seven of the differences are less than 1.5 %, and only one of those seven exceeds 2.0%. For incorrect scoring of current data, 20 differences exceed 1.5% and 10 exceed 2.0%. In contrast, of the 18 score levels that Freedle selectively chose to report in his Table 2 (see Table 3 above), 12 exceed 1.5%, 8 exceed 2.0%, and two are 3.0%.

Third, when incorrect scoring and correct scoring with the current data are compared, differences between the correct scoring and incorrect scoring columns for both Black and White examinees tend to increase as the fs^{-} score increases. Likewise, the Black–White differences tend to increase with increasing fs^{-}, which is not consistent with Freedle's findings reported in his Table 2 (our Table 3).

There are two potential reasons for the differences between Freedle's flawed findings and those reported in Table 8 for the incorrect scoring of current data. First, Freedle used an old SAT test edition that had not been screened for DIF.[5] Second, the current population of test takers is far more test savvy than the group studied by Freedle.

The new appropriately scored version of the hard test/total test relationship in Table 8 reveals that there is a small consistent difference between the expected scores on the hard half-test for Black examinees and White examinees: Black examinees tend to have slightly higher expected hard half-test scores. What would Freedle do with these differences? He would probably use them as a basis for his R-score.

Table 8

Correctly vs. Incorrectly Calculated Averages of Hard Test Scores by Total Test Scores for a November 2001 SAT Verbal Examination

Total test formula score (fs^-)	Rounded total test scale score	Incorrect scoring		Freedle linking		Correct scoring		Freedle linking	
		Freedle R-score R/(R+W)		Diff. in average %Max	Freedle "gain" score	Average item score (fs^-)		Diff. in average %Max	Freedle "gain" score
		Whites	Blacks	Blacks-Whites	Blacks	Whites	Blacks	Blacks-Whites	Blacks
1	25(0)	13.31%	14.01%	0.13%	0(0)	14.01%	14.05%	0.00%	0(0)
2	27(0)	14.31%	14.74%	–0.62%	–2(0)	14.74%	14.51%	–0.20%	0(0)
3	28(0)	14.56%	15.18%	–0.25%	–1(0)	15.18%	15.21%	0.00%	0(0)
4	30(0)	14.49%	15.06%	1.00%	2(0)	15.06%	15.96%	0.90%	2(0)
5	31(0)	14.31%	15.23%	1.28%	1(0)	15.23%	16.43%	1.20%	1(0)
6	32(0)	16.08%	16.54%	0.05%	1(0)	16.54%	16.68%	0.10%	1(0)
7	33(0)	16.15%	16.60%	1.23%	2(0)	16.60%	17.74%	1.10%	1(0)
8	33(0)	16.79%	17.23%	0.08%	0(0)	17.23%	17.36%	0.10%	0(0)
9	34(0)	18.13%	18.22%	0.18%	0(0)	18.22%	18.54%	0.30%	0(0)
10	35(0)	17.08%	17.42%	1.59%	1(0)	17.42%	19.10%	1.70%	1(0)
11	36(0)	18.62%	18.73%	0.15%	0(0)	18.73%	19.01%	0.30%	0(0)
12	36(0)	19.21%	19.31%	0.23%	0(0)	19.31%	19.43%	0.10%	0(0)
13	37(0)	19.74%	19.65%	0.62%	1(0)	19.65%	20.23%	0.60%	1(0)
14	38(0)	20.38%	20.19%	0.13%	0(0)	20.19%	20.40%	0.20%	0(0)
15	38(0)	20.74%	20.53%	1.26%	1(0)	20.53%	21.39%	0.90%	1(0)
16	39(0)	21.46%	21.32%	0.46%	0(0)	21.32%	21.60%	0.30%	0(0)
17	39(0)	22.03%	21.67%	0.23%	0(0)	21.67%	21.80%	0.10%	0(0)
18	40(0)	23.03%	22.52%	1.38%	1(0)	22.52%	23.59%	1.10%	1(0)
19	41(0)	23.82%	23.21%	1.26%	1(0)	23.21%	24.00%	0.80%	0(0)
20	41(0)	24.64%	23.75%	1.03%	1(0)	23.75%	24.32%	0.60%	1(0)
21	42(0)	25.33%	24.47%	1.08%	1(0)	24.47%	25.07%	0.60%	0(0)
22	42(0)	25.64%	24.79%	1.59%	1(0)	24.79%	25.90%	1.10%	1(0)
23	43(0)	27.10%	25.94%	0.57%	0(0)	25.94%	26.21%	0.30%	0(0)
24	43(0)	27.46%	26.34%	1.80%	2(0)	26.34%	27.80%	1.50%	1(0)

(Table continues)

Table 8 (continued)

Total test formula score (fs^{-})	Rounded total test scale score	Incorrect scoring		Freedle linking		Correct scoring		Freedle linking	
		Freedle R-score R/(R+W)		Diff. in average %Max	Freedle "gain" score	Average item score (fs^{-})		Diff. in average %Max	Freedle "gain" score
		Whites	Blacks	Blacks-Whites	Blacks	Whites	Blacks	Blacks-Whites	Blacks
25	44(0)	28.92%	27.09%	0.87%	1(0)	27.09%	27.70%	0.60%	0(0)
26	44(0)	28.92%	27.56%	0.49%	1(0)	27.56%	28.01%	0.40%	1(0)
27	45(0)	29.72%	30.82%	1.10%	0(0)	28.26%	28.98%	0.70%	0(0)
28	45(0)	31.00%	32.26%	1.26%	1(0)	29.39%	30.24%	0.80%	1(0)
29	46(0)	31.97%	32.82%	0.85%	0(0)	30.08%	30.41%	0.30%	0(0)
30	46(0)	32.95%	34.90%	1.95%	1(0)	30.61%	31.72%	1.10%	1(0)
31	47(0)	33.54%	34.90%	1.36%	0(0)	31.66%	32.50%	0.80%	0(0)
32	47(0)	34.92%	36.15%	1.23%	1(0)	32.70%	33.61%	0.90%	1(0)
33	48(0)	35.77%	36.92%	1.15%	0(0)	33.48%	33.99%	0.50%	0(0)
34	48(0)	37.05%	38.08%	1.03%	1(0)	34.50%	35.09%	0.60%	1(0)
35	49(0)	38.67%	40.33%	1.66%	1(0)	35.46%	36.43%	1.00%	0(0)
36	49(0)	38.95%	40.03%	1.08%	1(0)	36.33%	37.26%	0.90%	1(0)
37	50(0)	39.74%	41.31%	1.57%	1(0)	37.16%	38.26%	1.10%	0(0)
38	50(0)	40.95%	43.13%	2.18%	2(0)	38.28%	39.47%	1.20%	1(0)
39	51(0)	42.08%	44.92%	2.84%	1(0)	39.20%	41.13%	1.90%	1(0)
40	52(0)	44.00%	46.46%	2.46%	1(0)	40.39%	41.82%	1.40%	0(0)
41	52(0)	44.23%	45.62%	1.39%	1(0)	41.25%	42.54%	1.30%	1(0)
42	53(0)	45.33%	47.54%	2.21%	0(0)	42.52%	43.71%	1.20%	0(0)
43	53(0)	46.90%	48.72%	1.82%	1(0)	43.83%	45.22%	1.40%	1(0)
44	54(0)	48.15%	48.69%	0.54%	0(0)	44.91%	45.67%	0.80%	0(0)
45	54(0)	49.79%	52.26%	2.47%	2(0)	45.96%	47.65%	1.70%	1(0)
46	55(0)	50.33%	51.46%	1.13%	0(0)	47.29%	47.97%	0.70%	0(0)
47	55(0)	51.62%	53.31%	1.69%	1(0)	48.57%	49.87%	1.30%	1(0)
48	56(0)	52.85%	54.59%	1.74%	1(0)	49.94%	51.05%	1.10%	1(0)
49	57(0)	54.56%	57.28%	2.72%	1(0)	51.40%	53.19%	1.80%	0(0)
50	57(0)	56.67%	58.97%	2.30%	1(0)	52.53%	53.95%	1.40%	1(0)
51	58(0)	56.67%	59.64%	2.97%	1(0)	53.87%	55.96%	2.10%	0(0)
52	58(0)	58.77%	59.97%	1.20%	1(0)	55.50%	56.28%	0.80%	0(0)

(Table continues)

Table 8 (continued)

Total test formula score (fs^{-})	Rounded total test scale score	Incorrect scoring		Freedle linking		Correct scoring		Freedle linking	
		Freedle R-score R/(R+W)		Diff. in average %Max	Freedle "gain" score	Average item score (fs^{-})		Diff. in average %Max	Freedle "gain" score
		Whites	Blacks	Blacks-Whites	Blacks	Whites	Blacks	Blacks-Whites	Blacks
53	59(0)	60.00%	62.36%	2.36%	1(0)	57.07%	58.39%	1.30%	1(0)
54	60(0)	61.67%	62.49%	0.82%	0(0)	58.65%	59.20%	0.60%	0(0)
55	60(0)	63.38%	65.33%	1.95%	2(0)	59.70%	61.23%	1.50%	1(0)
56	61(0)	64.05%	66.56%	2.51%	1(0)	61.51%	63.31%	1.80%	1(0)
57	62(0)	65.49%	66.56%	1.07%	0(0)	63.14%	63.95%	0.80%	0(0)
58	62(0)	67.23%	68.21%	0.98%	1(0)	64.82%	65.63%	0.80%	0(0)
59	63(0)	68.97%	69.00%	0.03%	0(0)	66.63%	66.58%	–0.10%	0(0)

Let's review how he computed his R-score. Return to Table 3. The 4th column in Table 3 is a so-called score gain that is obtained by finding the closest predicted values on the hard half-test in columns 2 and 3, and using inverse regression[3] to obtain a score that appears to be on the 200–800 scale. For example, Freedle considers 16.1% in column 2 and 16.1% in column 3 to be equivalent. He goes into the regression backwards and obtains scores of 230 for Black examinees and 290 for White examinees. He takes the difference between these values, 60, which is in the 4th column. He then adds the 60 to scaled score of 230 for Black examinees to get a new score of 290 in column 5, which is the R-score. Finally, the 6th column is simply the difference in percentage correct on the half-test between Black examinees and White examinees.

What is wrong with this approach? Several things. He used a different group for each item to define his half-test regressions. As a consequence, the numbers in the table do not mean what he thinks they do. He used data that come from a test that was administered in March 1980, long before DIF was used to screen items on the SAT I. He used inverse regression for the White group to find a score on the total test that predicts a certain half-test score. He then used this White group inverse regression in conjunction with the Black examinee half test-scores to estimate his R-score for Black examinees.

Forget for a moment that this method of scoring is flawed. What would happen if we used Freedle's R-score technique with data that are correctly scored and properly pooled over items (i.e., the data in Table 8)? Look at the score level 51. When scored correctly, it yields the

largest difference (2.1%) in expected %Max in favor of Black examinees. According to Freedle's R-score procedure, we would look for the expected %Max for White examinees closest to the expected %Max of 55.96% for Black examinees. This would be the 55.50% associated with an fs^- of 52, which has a scale score of 584.55, which is about 6 points higher than the scale score of 578.35 associated with an fs^- of 51. Both 51 and 52 round to scale scores of 58(0). So in the last column of the correct Table 8, we find Freedle's "gain score" for correct scoring, which for a $fs^- = 51$ is 0(0).[6] In the case of $fs^- = 51$, the Freedle R-score based on correct scoring would be the same as the score based on the longer more reliable test. (A slightly more sophisticated version of inverse regression involving linear interpolation would yield a score of 59(0), one score unit higher than the one obtained by performance on the whole test.)

When scored using Freedle's $R/R+W$, we find a "gain score" of 1(0) for $fs^- = 51$) as can be seen in column 6 of Table 8.

What happens with the other large differences in average score on the hard half-test when correct scoring is used? At score level 39, the Freedle R-score would be 52(0) instead of 51(0). At score level 49, the R-score would remain 57(0). At score level 56, the R-score would be 62(0) instead of 61(0). At score level 45, the R-score would be 55(0) instead of 54(0). At score level 10, the R-score would be 36(0) instead of 35(0). For score level 55, the R-score would be 61(0) instead of 60(0). And at score level 24, we see a R-score of 44(0) instead of 43(0). Freedle's scoring (column 6) yields similar results with the exception that some score locations show 2(0) point differences.

Below fs^- scores of 13 on the total test, both the Black examinees and White examinees score below chance on the hard half-test (for fs^+, chance = 0.20 or 20%). If we limit our attention to the nonchance portion of the table above $fs^- = 13$, a pattern tends to emerge under correct scoring with every other score "gain" being 0 and the ones in between being 1. This pattern of no change alternating with the smallest change possible is markedly different from the results in Freedle's Table 2 (Table 3 in this report). When the hard half-test is scored properly and consistently with the total test, Freedle's findings almost evaporate.

3.4 Summative Evaluation of Freedle's Table 2

Freedle's Table 2 defines "percentage correct" as $R_i/(R_i+W_i)$ where the group used to define the "percentage correct" changes systematically from item to item in a way that depends on the difficulty and location of the item in the test. "Summing" these "percentages correct"

produced his Table 2 (Table 3 here). His table was constructed improperly. As a consequence, the numbers do not mean what he claims they mean.

The right portion of Table 8 shows what happens when the hard half-test is scored in a way that is mathematically consistent with the way in which the total test is scored. When the hard half-test is scored correctly and actual performance differences on the hard half-test are computed between a Black group that is the same across items and a White group that is the same across items, the score "gains" are very minor: 1(0) score unit or none at all. Even using Freedle's incorrect scoring procedure score gains were at most 2(0). Application of the R-score approach to a properly constructed table produces little or no effect on Black examinee scores, on average. The small effect associated with the dubious R-score that occurs with proper scoring is much more consistent with the DIF results cited at the beginning of Freedle (2003, p. 3) than are the large score "gains" generated by Freedle's incorrect score analysis.

4. The DIF/Difficulty Evidence: Then and Now

4.1 Differential Item Functioning (DIF)

Freedle used the results of DIF as the justification for his "bias" case, which led to his flawed Table 2 and his recommendation for use of the hard half-test in creating a Freedle-R score. In order to interpret DIF results, it is important to know what DIF is. Holland and Wainer (1993) is often cited as a reference book for DIF. In that book, Dorans and Holland (1993) make an important distinction between DIF and impact. Impact refers to a difference in performance between two intact groups, such as Blacks and Whites. Impact exists in test and item data because individuals differ with respect to the construct (e.g., mathematical proficiency, measured by the items) and intact groups of individuals differ with respect to their distributions of scores on measures of these constructs. For example, on a typical SAT-Math item, it is well-known that Asian Americans, as a group, score higher than Whites, as a group, and that men score higher than women. This is impact. Impact on one item is often consistent with impact on similar items.

In contrast to impact, which is affected by differences between groups in score distributions, DIF studies differences in how items function after differences in score distributions between groups have been removed statistically. Unlike impact, which confounds item differences with group differences, DIF examines item differences after an attempt has been made to control for group differences. DIF compares item performance of groups that are

comparable with respect to what the item is indeed to measure. While impact is a function of group differences as well as item differences, DIF is sensitive only to item differences.

Freedle (2003) recognized the need to adjust for overall performance differences and compare the comparable, so that differences in how an item functions across groups can be evaluated apart from group differences that affect performance comparisons on all items. He used DIF analysis instead of simply looking at impact, which is the difference in performance of intact groups of Whites and Blacks on test questions or items.

Rules have been set up for screening items on the bias of the size of the DIF measures. Freedle (2003, p. 3) noted that for the tests he examined, few items exhibited the kind of DIF that would attract attention in a screening process, not because the screening process was flawed, but because the amount of DIF was too small to be concerned about. In other words, Freedle himself noted that there was little or no DIF on the tests on which he concentrated. So at the item level, there was little DIF worth noting.

4.2 The Standardization DIF Procedure

There are many DIF procedures, as can be seen in Holland and Wainer (1993). The procedure used by Freedle was an unusual application of the standardization procedure, which was developed by Dorans and Kulick (1986) and which is described along with the more widely used Mantel-Haenzel procedure by Dorans and Holland (1993). The standardization approach can be thought of as a procedure for comparing two item-test regressions between two groups, a focal group (e.g., females, Blacks), and a reference group (e.g., males and Whites).

The standardization method uses the regression of item scores onto the total score. In other words, it computes the average item score (if the items are scored as 1 for correct and 0 otherwise, this average item score will fall between 0, an item that no one answers correctly, and 1, an item everyone answers correctly) for each possible score on the total test. We can compute this regression curve in both a reference group (e.g., Whites) and a focal group (e.g., Blacks), and then compute differences between these curves. The standardization procedure averages these differences across all the score levels to arrive at the number that describes the degree to which there are differences in how an item functions in the focal and reference groups. So the DIF statistic is a weighted average of differences between item-test regressions in a focal and reference group. The weights used are the proportion of individuals in the focal group at each

score level. For more details about standardization and Mantel-Haenszel, consult Dorans and Holland (1993).

4.3 DIF Results From Older SAT Forms

Although he found little DIF at the item level, Freedle reported a correlation of about 0.50 between the DIF statistic he used and the difficulty of the item. He found slight positive DIF for African-Americans on hard items and slight negative DIF on easy items. This means African-Americans did slightly better than a matched group of Whites on hard items and slightly worse than the matched Whites on easy items.

This finding was consistent with what had been found in the literature. Kulick and Hu (1989) reviewed the literature on the DIF/difficulty relationship, a literature that dated back into the 1970s. Kulick and Hu examined the relationship between DIF and difficulty for several groups on 765 verbal items from nine forms of the SAT administered between June 1986 and December 1987. For the Black/White DIF analysis, the correlation between DIF, as measured by the Mantel-Haenzel DIF (MHD-DIF) statistic, and item difficulty was 0.40, which was smaller in magnitude than the correlation, –0.51, between MHD-DIF and the standardized difference in omit rates between Black examinees and White examinees. For analogy items only, this DIF/difficulty correlation was 0.57 (DIF indices correlated –0.62 with differential omit rates). Burton and Burton (1993) reported a correlation between MHD-DIF and item difficulty from the Black/White DIF analysis conducted on 607 analogy items from the 1987–1988 verbal pretest pool to be 0.58.

These studies involved data from items that had not been prescreened for DIF. In addition, Freedle's analyses were performed on old versions of the SAT, as were the Kulick and Hu's and Burton and Burton's analyses. What is the situation currently? In particular, what has been the effect of DIF screening at the pretest stage, which was instituted in the late 1980s?

4.4 DIF Results From Current Tests

We analyzed 468 verbal items from six recent SAT I forms administered between October 2002 and May 2003. The correlations results for the verbal items are shown in Table 16. The variables are Difficulty (EqD), the Mantel-Hanzsel DIF (MHD-DIF), Standardized DIF, for formula scoring of the item (STDFS-DIF), Impact (IMP) and a measure of item correlation with the total score, (RBIS).

EqD is difficulty measured in the equated delta metric, the same measure used by previous authors, and is a transformed proportion correct. MHD-DIF is an estimated difference in deltas for the focal and reference groups after adjusting for ability differences and is used to screen every SAT I item for DIF before it appears on a test form that is used for operational scoring. STDFS-DIF summarizes differences in item-test regressions in which the item is score as 1 for a correct response, 0 for an incorrect response, and –1/(k–1) for a missing response, where k is the number of options for the item. Freedle used a different measure in which items were scored as correct, 1, or incorrect, 0, and examinees with missing responses were excluded from the analysis for that item. The measure used here is more appropriate because item scoring is consistent with total test scoring and all data are included in the analyses. (See sections 4 and 5 below for a detailed treatment of Freedle's scoring procedure.) IMPACT is the unadjusted difference in proportion correct between the focal group (e.g., African-American) and reference group (e.g., Whites). The RBIS is a measure of how correlated the item is with the total score in the total group. For more details on these measures, see Dorans and Holland (1993).

ETS uses the delta to measure difficulty, so that large numbers mean more difficult items. Many in the field use proportion correct as a measure of item difficulty, for which higher values are associated with easy items. Hence, the correlation between item difficulty and the separating power of the item (i.e., the RBIS) is usually positive with the proportion correct measure of difficulty, and negative when EqD measures difficulty. Whether EqD or proportion correct is used, easier items tend to have higher RBIS than hard items. This relationship between difficulty and RBIS has been cited as a statistical explanation for the relationship between difficulty and measures of DIF that presume that all items have the same level of separating power (Dorans, 1982).

4.5 Black/White Correlational Results

While the two DIF measures have a high correlation (0.97) in Table 9, their correlation with IMPACT is only 0.51 across all the items (above diagonal) and around 0.55 in the analogy items (below diagonal). Note also the correlation between IMPACT and RBIS is almost –0.65, meaning that group differences on the items tend to be smallest on the questions with lower measuring power. The correlations of 0.23 and 0.26 between IMPACT and EqD means that the impact tends to be smaller on the harder questions. Likewise, DIF is also smaller on the hardest questions as indicated by the 0.27 and 0.31 for the correlations between EqD and the MHD-DIF measure, which treats missing responses as incorrect, and the 0.19 and 0.22 for the correlation of

EqD with STDFS-DIF. Note that while the correlations between DIF and difficulty are slightly higher for the analogy items they are still quite small in magnitude.

Table 9

Correlations Among Black/White DIF Measures and Other Item Statistics

	EqD	MHD-DIF	STDFS-DIF	IMPACT	RBIS
EqD		0.27	0.19	0.23	–0.26
MHD-DIF	0.31		0.97	0.51	–0.09
STDFS-DIF	0.22	0.97		0.51	–0.07
IMPACT	0.26	0.54	0.56		–0.64
RBIS	–0.38	–0.13	–0.06	–0.63	

Note. Correlations among Black/White dif measures and other item statistics for 468 verbal items (above main diagonal) and 114 analogy items (below main diagonal) from six SAT I test forms administered from October 2002 to June 2003.

What do correlations of a 0.27 or a 0.19 mean in practical terms? Often the correlation is squared to obtain a measure of how much variation in one variable is shared with another variable. Correlations of about 0.5 indicate that the two variables share 25% of their variance. This degree of association is noteworthy. Freedle (2003), Kulick and Hu (1989), and Burton and Burton (1993) report correlations in the neighborhood of 0.5.

Proportions of shared variance that are less than 10% rarely excite much interest. A correlation of 0.27 means that 7% of the variation in one variable can be accounted for by knowledge of the other variable, leaving the remaining 93% unexplained. A 0.31 correlation is near 10%. Correlations of 0.27 and 0.31 are not strong correlations. Correlations of 0.19 and 0.22 are even weaker with shared variances of less than 5%. None of these correlations appear large enough to generate a modest amount of interest, let alone be used as a justification for bias that requires reporting special group-specific scores for the SAT.

DIF screening seems to have successfully reduced the degree of correlation between DIF and difficulty that served as the impetus for Freedle's provocative claims and recommendation for the reporting of Freedle R-scores for non-White groups. Direct evidence for this claim of reduction in the DIF/difficulty relationship can be found in the DIF analyses performed from October 2002 to June 2003 on 614 verbal pretest items. The DIF/difficulty relationship in these unscreened items are represented in the results for Black/White DIF data, presented in Table 10.

Table 10

Correlations Among Black/White DIF Measures and Other Item Statistics for 614 Verbal Pretest Items (Above Diagonal) and 121 Analogy Pretest Items (Below Diagonal) From Six SAT I Test Forms Administered From October 2002 to June 2003

	EqD	MHD-DIF	STDFS-DIF	IMPACT	RBIS
EqD		0.39	0.28	0.37	–0.38
MHD-DIF	0.48		0.95	0.62	–0.17
STDFS-DIF	0.36	0.95		0.63	–0.15
IMPACT	0.37	0.77	0.76		–0.67
RBIS	–0.38	–0.26	–0.22	–0.53	

The DIF/difficulty relationship, while not at the level of that reported with data from the old test by Kulick and Hu, and Burton and Burton, is noticeably higher than that reported in Table 9. There appears to be a DIF/difficulty relationship that merits some investigation. There is a strong possibility that this relationship is a statistical artifact due to the fact that large group differences in test score distributions are not completely dealt with by the standardization, Mantel-Haenszel or, perhaps, any observed-score DIF procedure, and that impact remains in the data even after adjusting for total score differences. Since impact is positively related to difficulty, any DIF remaining in the impact may also be related to difficulty. There is a methodological issue here that needs to be addressed before we speculate about why the relationship exists.

Regardless of whether or not there is a methodological issue, there is little evidence in the DIF data to support Freedle's claim of bias on SAT I test forms. In addition, there is definitely no evidence to support his recommendation to supplement SAT scores with scores based on the hardest part of the test or scores that depend on group membership. Over a decade of DIF screening on the SAT appears to have muted the relationship between DIF and difficulty that served as his dubious starting point. In short, Freedle's DIF results have questionable relevance because the test changed and because examinees are much more test savvy than they were nearly 25 years ago.

5. Conclusions and Future Research

5.1 Conclusions

Freedle's analyses are flawed on many levels.

- Freedle's hard half-test is a flawed idea. It is less reliable, especially for less proficient examinees. Less reliable means less consistent ordering of examinees on the basis of their proficiency and more lottery-like noise in each score.
- Freedle's scoring at the item level excluded examinees who did not respond to the question, and his "averages" are therefore based on an indeterminate group.
- Freedle used improperly conducted analyses of DIF on obsolete SAT tests to rationalize his proposed R-score. The particular test edition he used was administered prior to the time that test disclosure altered the test-taking practices of examinees who took the SAT and other high stakes tests.

5.2 Related Research and Future Research

Elsewhere, we demonstrate that Freedle's use of inverse regression violates basic tenets of sound score linking (Dorans & Zeller, 2004a), and that there are sounder ways to assess whether the scores on a hard half-test are comparable to scores on the full test (Dorans & Zeller, 2004b). There remain other avenues for research that ought to be explored as well.

It is clear from Figure 1 that Black examinees and White examinees, on average, achieve comparable formula scores in different ways. At lower formula scores, there are only small differences between Black examinees and White examinees in proportion choosing the correct response, larger differences in proportions selecting an incorrect response (White examinees more than Black examinees), and larger differences in nonresponse (Black examinees not responding as often as White examinees).

Part of this difference might be attributed to the fact that Black examinees complete fewer questions at the end of test sections relative to White examinees with comparable total scores (Schmitt, Dorans, & Holland, 1993). Questions at the end of the test tend to be harder questions. If one omits or fails to reach a hard question, he or she will get a higher score than if the question is attempted and results in an incorrect answer. Failure to reach the harder questions at the end of sections may boost the scores of Black examinees over matched White examinees.

Future research should examine this differential speededness hypothesis, as well as other hypotheses, including Freedle's proposed explanations, and explanations that maintain that the effect is a statistical artifact. Keep in mind that the effect to be studied, however, is a very small effect, once the half-test is scored properly. Schmitt et. al (1993) provide detailed advice about how to carry out observational and experimental studies that try to tease out the sources of these small effects.

References

Burton, E., & Burton, N. W. (1993). The effect of item screening on test scores and test characteristics. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 321–335), Hillsdale, NJ: Lawrence Erlbaum Associates.

Dorans, N. J. (1982). *Technical review of item fairness studies: 1975–1979* (SR-82-90). Princeton, NJ: ETS.

Dorans, N. J. (2002). Recentering the SAT score distributions: How and why. *Journal of Educational Measurement, 39*(1)*,* 59–84.

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

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 35–66), Hillsdale, NJ: Lawrence Erlbaum Associates.

Dorans, N. J., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. *Journal of Educational Measurement, 23,* 355–368.

Dorans, N. J., & Zeller, K. (2004a). *Inverse regression is an inappropriate score linking procedure: A critique of a scoring procedure that allegedly reduces ethnic differences.* Manuscript in preparation.

Dorans, N. J., & Zeller, K. (2004b). *Using score equity assessment to evaluate the equatability of the hardest half of a test to the total test*. Manuscript in preparation.

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. (1995). *Test equating: Methods and practices.* New York: Springer-Verlag.

Kulick, E., & Hu, P. G. (1989). *Examining the relationship between differential item functioning and item difficulty.* (College Board Report No. 89-5; ETS RR-89-18). New York: College Entrance Examination Board.

Lawrence, I. M., Rigol, G. W., Van Essen, T., & Jackson, C. A. (1993). *A historical perspective on the SAT® 1926–2001* (College Board Research Report No. 03-03, ETS RR-03-10). New York: The College Board.

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.

Zieky, M. (1993). Practical questions in the use of DIF statistics in test development. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 337–347). Hillsdale, NJ: Erlbaum Associates.

Notes

[1] It is not clear why Freedle used a 1980 form. Freedle's analyses were based on old SAT forms that were built to different specifications than the current test. (See Lawrence, Rigol, Van Essen, & Jackson, 2003, for a review of the content changes in the SAT during the 20th century.) In addition, students taking a 1980 form were not exposed to the growth in test preparation that followed the test disclosure legislation of 1980 that might have improved familiarity with question formats and scoring instructions.

[2] 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. From the data in Table 2, for example, if we knew a person's score on the total test we could predict their score on the hard test. 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.

[3] If k represents the number of options in the multiple-choice question, then the fs^{-} rule assigns the quantity $-1/(k-1)$ to an incorrect answer. In this way, if a person responded randomly to the entire test, we would expect the person to receive a total formula score of zero. The logically equivalent fs^{+} rule assigns the quantity $1/k$ to an unanswered item. Both scoring rules assign the lowest score to an incorrect response, a slightly higher score to a non-response, and the highest score to a correct response.

[4] We put scores on a 20(0) to 80(0) scale to emphasize what is often lost in interpreting SAT scores: the third zero means nothing, so a 10 point increase on the SAT scales is only a one unit change in score value. In other words, because there are no reported scores between 500 and 510, the change is actually only one score unit. This is more obvious when the third "0" is dropped. As a compromise, we will use 20(0) to 80(0) to indicate the 200–800 point scale.

[5] Freedle's analysis involved data from tests that were not screened for DIF at the pretest stage. Pretesting of items before they are used in actual test editions and count towards scores is routinely performed on the SAT I to ensure that items perform well statistically. One of the performance criteria is that they are acceptable from a DIF perspective (Zieky, 1993).

[6] As noted earlier, we are using (0) to emphasize the fact that the change in score units occurs with change in the tens place, not the units place, and to emphasize the small nature of a 1(0) point gain as a change of a single score unit.