Unfair Treatment? The Case of Freedle, the SAT, and the Standardization Approach to Differential Item Functioning

MARIA VERONICA SANTELICES
Pontificia Universidad Católica de Chile

MARK WILSON
University of California, Berkeley

In 2003, the Harvard Educational Review *published a controversial article by Roy Freedle that claimed bias against African American students in the SAT college admissions test. Freedle's work stimulated national media attention and faced an onslaught of criticism from experts at the Educational Testing Service (ETS), the agency responsible for the development of the SAT. In this article, Maria Veronica Santelices and Mark Wilson take the debate one step further with new research exploring differential item functioning in the SAT. By replicating Freedle's methodology with a more recent SAT dataset and by addressing some of the technical criticisms from ETS, Santelices and Wilson confirm that SAT items do function differently for the African American and White subgroups in the verbal test and argue that the testing industry has an obligation to study this phenomenon.*

Because the SAT I is the most widely taken college admissions test—with more than 1.5 million students taking it every year (College Board, 2008)—its fairness to all subgroups in the population is an issue that concerns us all. Test validity and potential test or item bias against disadvantaged students affects access to postsecondary education and thus has grave long-term consequences.

The validity of admissions tests as a selection tool for higher education is, technically, judged by two main criteria: the test fairness to all subgroups of the population and the degree to which test scores predict college outcomes for different subgroups (differential prediction). Test developers utilize vari-

Harvard Educational Review Vol. 80 No. 1 Spring 2010

ous concepts and statistical methods designed to ensure that tests are fair to all subgroups. Differential item functioning (DIF) is one of those concepts and refers to a psychometric difference in how an item functions for two subgroups of examinees matched by proficiency with respect to the construct being measured (Dorans & Holland, 1992). DIF analyses tell us about test performance differences that are unrelated to the construct that the test purports to measure and that are the result of nuisance factors that should ideally not influence test results. Several prominent resources for testing professionals, such as the Standards for Psychological and Educational Testing (AERA, APA, & NCME, 1999), the fourth edition of *Educational Measurement* (Brennan, 2006), and the Code for Fair Testing Practices in Education (Joint Committee on Testing Practices, 2004) cite DIF analyses and methodologies as critical for assessing test fairness and the validity of test scores.

It is important to distinguish between differential impact and DIF. *Differential impact* refers to a difference in performance between two groups, such as African Americans and Whites; it exists in test and item data because individuals differ with respect to the construct tested (e.g., mathematical proficiency) and groups of individuals differ with respect to their distributions of scores on measures of these constructs. DIF studies, however, look at how items function after differences in score distributions between groups have been statistically removed. The remaining differences indicate that the items function differently for both groups. Typically, the groups examined are derived from classifications such as gender, race, ethnicity, or socioeconomic status. The performance of the group of interest (focal group) on a given test item is compared to that of a reference or comparison group. White examinees are often used as the reference group, while minority students are often the focal groups. This is done because White examinees are usually the majority in terms of numbers and therefore the results are easier to interpret. The goal of test developers is to create assessments with no DIF between groups, since, by definition, these differences in performance are irrelevant to the construct measured and render test results invalid. Because of the way it is calculated, DIF is an item-bias indicator, not a test-bias indicator, and judges the functioning of each item against the functioning of the test as a whole. Thus, DIF is an internal-reference, as opposed to an external-reference, criterion to study bias and is significantly shaped by the functioning of items in the largest subgroup of examinees—in this case, White students.[1]

DIF could have numerous causes, such as language that different subgroups might interpret differently or content to which subgroups might have differential exposure (e.g., culturally specific content). However, the DIF methodologies most often used to study the SAT are descriptive and cannot identify the reasons for differential item performance. If differences in favor of a given group are found in certain items, then further analyses are required to determine the cause of the DIF. These investigations might employ qualitative methodology that explores cognitive processes students undergo when read-

ing and answering test questions (e.g., think-alouds) or quantitative methodology that associates differences in item functioning with observed variables such as item difficulty or item position (e.g., correlations or explanatory item response theory).

One option is to relate DIF results to possible explanatory variables. This is the approach Roy Freedle used in his 2003 piece in the *Harvard Educational Review*, which brought the relationship between item difficulty and DIF to the nation's attention when an article in the *Atlantic Monthly* highlighted Freedle's findings (Mathews, 2003). Freedle's work asserted that many of the more difficult SAT items exhibited DIF benefiting African American students, while easier SAT items showed DIF favoring White students. Freedle, a cognitive psychologist who worked at the Educational Testing Service (ETS) for more than thirty years, hypothesized that the relationship between item difficulty and DIF estimates was explained by cultural familiarity and semantic ambiguity. "Easy" verbal items, he reasoned, tap into a more culturally specific content; therefore he hypothesized that these were perceived differently, depending on one's particular cultural and socioeconomic background. "Hard" verbal items, his hypothesis continued, often involve rarely used words that are less likely to have differences in interpretation across ethnic communities because these are only familiar to those with higher levels of education.

The empirical evidence on differential item functioning presented in Freedle's article was obtained using the "standardization approach." This statistical procedure, developed by ETS researchers Dorans and Kulick (1983), is based on the computation of expected scores for the focal and reference groups at each score level and allows one to control for differences in the subpopulation ability.[2] Dorans (Dorans, 2004; Dorans & Zeller, 2004a) strongly criticized Freedle for an erroneous application of the standardization approach. Dorans's criticism focused on Freedle's use of incorrectly defined groups at each score level and on the use of proportion-correct scores rather than formula scores.[3] He further criticized Freedle for using data that preceded the 1980 implementation of a DIF and bias review process for all ETS items (Ramsey, 1993; Zieky, 1993).

The research we present here aims both to replicate Freedle's work with a more current SAT dataset and to address Dorans's methodological concerns. Specifically, we employ the standardization approach used by Freedle and then compare these results to those obtained when Freedle's method is adjusted to respond to the criticisms of Dorans and his colleagues: when (1) the denominator used in the proportion-correct formula of the standardization p-difference index is calculated based on the recommendations of the ETS researchers, and (2) total scores take into consideration the possibility of students getting a right response as a result of guessing (formula score is used in place of the proportion-correct score). The item-level information we use in this study was collected in 1994 and 1999, well after ETS began using DIF sensitivity reviews.

Although modifications were made to the SAT in 2005, these changes do not affect the generalizability and relevance of our results. College Board authors have acknowledged that the 2005 revisions to the verbal and math sections were minor and that the main revision was the inclusion of a writing section (Mattern, Patterson, Shaw, Kobrin, & Barbuti, 2008). Changes to the verbal section (now known as critical reading) included the elimination of analogies and the addition of shorter reading passages. The math section was also minimally modified; quantitative comparisons were removed and content from the third year of high school was added.

Regardless of content modifications in recent years, identifying bias in past SAT tests is still critically important, given the role of the SAT in admission to selective institutions of higher education and the role these institutions play in dispensing rewards and benefits to members of our society. Fairness is a critical part of test validity, and tests should be valid when individual consequences are attached to them. Admissions decisions that rely heavily on a test that is not a valid measure for all subgroups of the population raise critical questions of fairness.

Research on Differential Item Functioning in the SAT for Minority Group Examinees

During the 1980s and early 1990s, more than twenty studies researched differential item functioning in ETS tests using either the standardization (Dorans & Kulick, 1986) and/or the Mantel-Haenszel procedures (Holland & Thayer, 1988). While some of these studies were exploratory analyses of the statistical significance of DIF indices and their relationship with item characteristics (Lawrence, Curley, & McHale, 1988; Schmitt & Bleistein, 1987), others followed a confirmatory (hypothesis-testing) line of inquiry in which a few factors, based on previous research findings, were hypothesized as possible sources of DIF (Carlton & Harris, 1992; Freedle & Kostin, 1991; Kulick & Hu, 1989).

DIF studies have usually compared the performance of White students to that of Latinos, Asian Americans, and African Americans (Schmitt & Dorans, 1988). As a way of isolating results from the effect of variations in language proficiency, most of the studies have excluded students who report speaking a language different from English as their "best language" in the Student Descriptive Questionnaire (SDQ)[4] (Schmitt & Dorans, 1988). Research has focused primarily on the SAT verbal test, particularly on the now-obsolete analogies section. Although the SAT math test has been studied for DIF, the DIF estimates from those investigations are smaller and less systematic and interpretable than those from the verbal test (Kulick & Hu, 1989; Schmitt & Dorans, 1988; Zwick, 2002).

One well-established finding regarding the SAT verbal test is the superior performance of African American and Latino examinees, compared to matched White examinees, on items with content that is especially relevant to African

American and Latino examinees (Carlton & Harris, 1992; O'Neill & McPeek, 1993; Schmitt & Dorans, 1988).[5] This content-related DIF appears consistently on verbal items in two item formats: sentence completion items and items based on reading passages. Another frequent finding is that when verbal items are classified by content, African American examinees are found to perform better than matched White examinees on items dealing with human relationships and worse than matched White examinees in analogies that contain science content (Freedle & Kostin, 1991; Schmitt & Bleistein, 1987).[6] Researchers have also systematically encountered DIF favoring White examinees in analogy items (Carlton & Harris, 1992; Schmitt & Bleistein, 1987).

Research on differential item functioning has also examined the relationship between DIF statistics and item characteristics such as item type, subject matter content, position of item, and word abstractness, among others (Carlton & Harris, 1992; Schmitt & Bleistein, 1987). The relationship between DIF and item difficulty has been a key focus of research. Ideally this relationship would be close to zero. Kulick and Hu (1989) found from an African American/White DIF analysis of 765 verbal items that the correlation between DIF, as measured by the Mantel-Haenszel DIF (MHD-DIF) statistic, and item difficulty was 0.40. For analogy items only, the DIF/difficulty correlation was 0.57. Burton and Burton (1993) reported a correlation between MHD-DIF and item difficulty from a Black/White DIF analysis conducted on 607 analogy items from the 1987–1988 verbal pretest item pool to be 0.58. Freedle (2003) found a correlation of about 0.50 between the DIF statistics he used and the difficulty of the items. Dorans and Zeller (2004a) also reported a correlation of 0.27 between equated deltas[7] and MHD-DIF and a correlation of 0.19 between equated deltas and the standardized formula score statistics for the African American/ White comparison of verbal items. They also found correlations of 0.31 and 0.22 for analogy items.

Since item difficulty is confounded with item order, subject matter, and word abstractness, among other factors, the interpretation of results regarding DIF and item difficulty is debatable.[8] ETS researchers have acknowledged that potential interactions among these variables do not allow for the complete isolation of the factors causing DIF (Dorans & Lawrence, 1987; Kulick & Hu, 1989).

Nevertheless, ETS researchers have offered explanations for this DIF phenomenon, often following one or more of three main lines of argument: (1) differential speededness (Schmitt & Bleistein, 1987); (2) differential guessing strategies among examinees of different racial/ethnic groups (Kulick & Hu, 1989); and/or (3) methodological issues (Dorans, 2004; Dorans & Zeller, 2004a).

In contrast, Freedle (2003) emphasized the role of linguistic and cultural differences among groups as an explanation for the relationship between DIF and item difficulty. Freedle's preference for the cultural explanation is an extension of previous results in which he and his coauthors (Freedle & Kos-

tin, 1988, 1991; Freedle, Kostin, & Schwartz, 1987) tested the ETS researchers' hypotheses regarding the role of speed and could not find evidence supporting it. Further, in a study in which item difficulty was not confounded with item position (Freedle et al., 1987), African American examinees still performed differentially better on the hard analogy items and differentially worse on the easy analogy items. In another study in which speeded and nonspeeded[9] items were analyzed, Freedle and Kostin (1991) found that science content, item difficulty, and analogy stems with part/whole relationship explained equally well the DIF statistics obtained through the standardization approach. In addition, they compared low-scoring White students to low-scoring African American students and high-scoring White students to high-scoring African American students. The same DIF pattern found in the whole sample was found in each of these subgroup comparisons, suggesting that the relationship between item position and DIF values is independent of a student's ability level.

Freedle (2003) asserted that the linguistic and cultural difference was present in all multiple choice tests regardless of the content area being tested. His claim was based on research on math SAT items conducted by other ETS researchers (Kulick & Hu, 1989) as well as on his own analyses of the multiple choice section of two Advanced Placement tests (U.S. history and biology) and math SAT items. The phenomenon was consistently present when comparing item performance between White and African American students, as well as between Hispanic and Asian students whose preferred language was English, and it became stronger when students whose preferred language was other than English were included in the analyses. The DIF pattern held even when item performance was compared between White students who reported English as their preferred language and White students who reported a preferred language different from English. These findings, in addition to the work of Diaz-Guerrero and Szalay (1991), led Freedle (2003) to state that the source of the phenomenon was not ethnicity but any index "that identifies a group as sharing a persistent environment that differs from the White majority English speakers" (p. 19).

As a result, Freedle proposed an alternate score to correct for the unfairness that DIF generates: the R-SAT. The R-SAT score measures student performance on the hard half of the SAT test, be it the verbal or math parts of the test. According to Freedle, the R-SAT would reduce the mean score differences between white and minority test-takers by a third and would be a more accurate indicator of the academic skills of minority students.

Researchers at ETS (Bridgeman & Burton, 2005; Dorans, 2004; Dorans & Zeller, 2004a, 2004b) have responded to Freedle's work by attributing his findings to statistical artifacts, technical problems in the implementation of the methodology he used to identify DIF (the standardization approach), and the effects of student guessing. In their view, issues of improper scoring[10] and score linking, in addition to the use of obsolete data,[11] render Freedle's findings insignificant. They also present analyses showing that the R-SAT

score is invalid[12] and unreliable and would have no effect on reducing group differences.

Research Questions

The research presented here seeks to counter the ETS criticisms by asking the following questions:

1. Is the phenomenon observed by Freedle—the relationship between item difficulty and DIF estimates for the White/African American comparison—present when the standardization approach advocated by ETS is implemented in datasets using SAT items that underwent ETS's DIF screening process?
2. Under these conditions, does the Freedle phenomenon hold across different ethnic groups?
3. Is the relationship between item difficulty and DIF estimates still present when the formula score is used in place of the proportion-correct score?

The current investigation does not address causal questions regarding the reasons for DIF. That question is left for future research since it requires a different methodology and is not justified until the replicability and generalizability of Freedle`s findings are established. Thus, we cannot judge the validity of the hypothesis advanced by Freedle (2003) and Diaz-Guerrero and Szalay (1991) about cultural and linguistic differences among the White, African American, and Hispanic groups.

Methods

The present study focuses on the methodological explanations offered by ETS researchers for the systematic relationship between item difficulty and DIF estimates in Freedle's findings. To do this, we mirrored Freedle's and ETS researchers' applications of observed-score methods in prior research (Dorans & Kulick, 1983; Dorans & Zeller, 2004a; Freedle, 2003; Freedle & Kostin, 1991).[13] We measured item difficulty using p-values—the proportion of examinees that answered an item correctly, which can range between 0.0 and 1.0. Higher p-values indicate that a greater proportion of examinees responded to the item correctly; a large p-value thus indicates an item that is experienced as relatively easy among the entire test population.

We studied DIF using the standardization approach, which compares empirical item characteristic curves using the total test score as an estimate of ability. The expected item score for each possible full-test score is computed for both a reference and a focal group, and then the differences between these curves are examined. The *standardization approach* was introduced by Dorans and Kulick (1983) and controls for differences in both subpopulation ability and item quality.[14] These numerous nonparametric item-test curves are sum-

marized via a numeric index in which the differences between item-test curves are weighted by the proportion of individuals in the focal group at each score level. The spirit of this method is to compare how an item functions in two groups of interest (i.e, Whites and African Americans) considering only examinees of the same ability level. This is done studying one item at a time. Two versions of the standardization approach can be found in the literature: the standardized p-difference (or STD P-DIF) and the standardized formula-score difference (or STD FS-DIF). Both are presented below.

Standardized P-Difference

The following index is called the standardized p-difference because the original application of the standardization methodology defined the expected item score in terms of proportion correct at each score level. It ranges from –1 to +1.

The standardized p-difference index for test items is given by:

$$STD\,P-DIF=\sum_{m=1}^{m} w_m (P_{fm}-P_{rm}), \qquad (1)$$

where

m = 1, 2, . . . , m is the full-test score level,
P_{fm} = proportion of the focal group answering the item correctly,
P_{rm} = proportion of the reference group answering the item correctly,
w_m = relative frequency of the focal group within score level m.

Although w_m could represent the relative frequency of any group of interest at score level m, the relative frequency of examinees in the focal group has been most commonly used because it gives the greatest weight to the differences in P_{fm} and P_{rm} at the score levels most frequently attained by the focal group under study. This index was originally designed for binary-scored tests. (See Clauser and Mazor (1998) for an overview of DIF methods for polytomous-scored tests and Zwick, Donoghue, and Grima (1993) for an application of the Mantel-Haenszel procedure to polytomous items.)

— Different Versions of STD P-DIF

Much of the methodological controversy around Freedle's work centers on the definition of the proportion-correct score. Generally a proportion-correct score is the ratio of the number of correct responses to a possible total:

$$P = R/T, \qquad (2)$$

where

P = proportion of items answered correctly by a given test-taker,
R = number of correct responses,
T = total number of items.

Freedle (2003), in the appendix to his paper, identifies three formulas for calculating a percentage-correct score. All three formulas use the number of

correct responses in the numerator (*R*, or number of right responses) but differ in the denominator. The denominator can be defined in several ways:

$$\begin{aligned} D_1 &= R+W, \\ D_2 &= R+W+O, \text{ or} \\ D_3 &= R+W+O+NR, \end{aligned} \qquad (3)$$

where

R = number of correct answers,
W = number of wrong answers,
O = number of omitted items,
NR = number of items not reached.

Note that omits are defined as blank items that occur among the correct and incorrect answers (i.e., before the not-reached items). Not-reached items are defined as omits occurring at the end of a test section.

Freedle presented tables with statistics based on different definitions of the denominator. For one of the main tables (table 2, p. 10), he used D_1, but for others, he used D_3. His choice of D_1 over D_3 for table 2 was strongly criticized by Dorans because the exclusion of omits and not-reached items results in differences in the group of students answering particular items, so that the group used to study a given question may vary in a systematic way related to the difficulty of the question and the proficiency of the examinees taking it. Dorans (2004) argues that "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" (p. 64).

Freedle's use of D_1, however, was not completely arbitrary. As he explains in the appendix, he based his choice on an old argument with Kulick and Hu (1989), who claimed that, since Whites actually have a higher item-omission rate than African Americans on the SAT, the inclusion of omits in the denominator artificially depresses the performance of Whites on the "hard" items relative to African Americans. They suggest that if Freedle and his coauthors had, in earlier work (Freedle & Kostin, 1988; Freedle et al., 1987), used the R+W denominator formula, they would not have found any systematic bias effects. Thus, by choosing to use D_1 over D_3 in table 2, Freedle was trying to show that Kulick and Hu were wrong in attributing the systematic relationship between item difficulty and DIF estimates to the denominator formula used.

Standardized Formula-Score Difference

For the computation of the standardized p-difference, the definition of total score was always based on the proportion of items answered correctly. However, the SAT is a formula-score test. Formula scoring takes into account the potential random guessing that might occur in multiple choice tests and lowers the score accordingly to make it comparable to scores obtained on tests with open-ended or fill-in-the-blank items. Formula scoring assumes that an examinee either knows the answer to a question or guesses at random from

among the choices (Frary, 1988). With a five-choice item, about one guess in five would be expected to be correct. That is, when guessing randomly, one item out of five would be right and four would be wrong. In order to reduce the final score by the gain expected from random guessing, one point should be discounted for every four items guessed at but answered incorrectly. This is the rationale behind the following formula:

$$FS = R - \frac{W}{(k-1)}, \qquad (4)$$

where

FS = formula score on the test for a given test-taker,
R = number of items answered right,
W = number of items answered wrong,
k = number of choices per item.

Formula scoring is designed to adjust scores for the gains due to random guessing. That is why the instructions for the SAT advise students to avoid completely blind guessing. Under this type of scoring, the expected item performance is computed by giving one point for each correct response, giving no point for omits, and discounting a fraction of a point for each incorrect response. The fraction to be discounted is given by the number of alternatives provided in the item.

Dorans and Holland (1992) recommend the standardized formula-score difference for assessing DIF in formula-scored tests like the SAT. Dorans and Zeller (2004a) criticized Freedle's use of the standardized p-difference and endorsed the use of the standardized formula-score difference as "more appropriate because item scoring is consistent with total test scoring and all data are included in the analyses" (p. 25).

In order to implement the STD FS-DIF, equation 1, which is used to estimate the standardized p-difference index for test items, needs to be redefined in terms of expected item performance. The equation below shows the expected item performance in the focal group, *f*, at score level *m:*

$$E_{fm} = \left[(R_{fm} * 1) + \left(W_{fm} * \frac{-1}{(k-1)} \right) + (O_{fm} * 0) \right] \Big/ N_{fm}, \qquad (5)$$

where

E_{fm} = expected item performance of a given item in the focal group, given overall score level m,
R_{fm} = number of items answered correctly in the focal group, given score level m,
W_{fm} = number of wrong responses in the focal group, given score level m,
k = number of choices per item,
O_{fm} = number of omits in the focal group, given score level m,
N_{fm} = sum of R_{fm}, W_{fm}, and O_{fm}.

For simplicity, this equation treats omits and not-responses as the same.

Correspondingly, the expected item performance in the reference group, *r*, at score level *m* is

$$E_{rm} = \left[(R_{rm} * 1) + \left(W_{rm} * \frac{-1}{(k-1)} \right) + (O_{rm} * 0) \right] \Big/ N_{rm}, \quad (6)$$

where terms are as defined for equation 5 but now refer to the reference group.

The redefined formula for the standardized formula-score difference is provided in equation 7 and incorporates equations 5 and 6.

$$STD\,FS - DIF = \sum_m w_m \left(\left[(R_{fm} * 1) + \left(W_{fm} * \frac{-1}{(k-1)} \right) + (O_{fm} * 0) \right] \Big/ N_{fm} - \left[(R_{rm} * 1) + \left(W_{rm} * \frac{-1}{(k-1)} \right) + (O_{rm} * 0) \right] \Big/ N_{rm} \right), \quad (7)$$

The standardized formula-score difference ranges from $-k/(k-1)$ to $k/(k+1)$. The STD FS-DIF and the STD P-DIF are more likely to diverge when items are difficult and omitting becomes a dominant factor (Dorans & Holland, 1992). None of the standardization statistics is sensitive to nonuniform DIF, which means that we cannot infer from them whether the severity of DIF is different for lower- and higher-ability test-takers.[15]

DIF Statistics Effect Size

Dorans and Holland (1992) provide the following guidelines regarding DIF effect size identified by these methods. These guidelines apply to the p-difference index:

- Values below |0.05| should be considered negligible.
- Values between |0.10| and |0.05| are considered medium sized and should be inspected to ensure that no possible effect is overlooked.
- Items with DIF values outside the [–0.1, +0.1] range are considered serious and should be examined very carefully.[16]

Data Source and Sample

We conducted this study using data provided by the College Board. Specifically, we analyzed the response patterns of all California students from public schools who took two specific SAT forms used in 1994 (QI and DX) and two specific SAT forms used in 1999 (IZ and VD).[17] The dataset from the College Board also contained the SAT verbal and math scores as well as students' responses to the Student Data Questionnaire (forty-three questions), which provides self-reported demographic and academic information such as parents' education, family income, and high school grade-point average. Most of the students in the dataset were high school juniors. To minimize the effects of differences in language proficiency in the population studied (Schmitt &

Dorans, 1988), we further limited the sample only to high school juniors who reported English as their "best language."[18]

The SAT item-level information is not frequently available to independent researchers who are not affiliated with ETS or the College Board, and we obtained the data under very special circumstances, namely because Richard Atkinson, the president of the University of California system at the time, was interested in studying the SAT in depth.[19] It took about two years to obtain the data.

Statistical Analyses and Results

Here, we present two sets of results. The first shows the findings obtained when applying Freedle's methodology to the SAT Form IZ, which underwent the screening for DIF items at pretest and was administered in 1999. Since the test had already been screened for DIF items, we did not expect to find items with large or medium DIF effect size. We programmed the standardization algorithm using SAS statistical software (package version 9.0) and used it to analyze differences in the responses provided by African American and White examinees. We first used the most controversial variant of the standardization p-difference used by Freedle: the formula using denominator D_1, which considers only right and wrong responses and excludes omits and not-reached items. The focus of this analysis was the comparison of these results to (1) those obtained with denominator D_3, which considers all of the items, and (2) those obtained when the formula score is used in place of the proportion-correct score, as Dorans and Zeller (2004a) suggest. We analyzed the relationship between item difficulty (p-values) and DIF estimates using correlation analysis and contrasts among item groups of different difficulty levels.

The total scoring we used in the calculation of the standardization indices is the formula score used by ETS for the SAT: correct responses receive a value of 1, nonresponses (omits or not reached) are scored as zero, and wrong answers are considered as $-1/(k-1)$ where k is the number of response options. We added scores over all items and converted to scaled scores using that year's corresponding table conversion from ETS. When we analyzed items from either the verbal or math SAT tests, we matched students on the total SAT formula score for that test only.

The second set of analyses explores the generalizability of the first set of results across ethnic groups and test forms. Freedle claims that the linguistic ambiguity explanation holds across languages and ethnic groups. To shed light on whether the patterns do occur across groups, we implemented both standardized p-difference and standardized formula-score difference indices to compare item performance first between White and African American students and then between White and Hispanic students on the four SAT forms (DX, QI, IZ, VD). The Hispanic group included students who reported in their SDQs being Latin American, Mexican, or Puerto Rican.

Differential Item Functioning in the SAT Form IZ

We chose Form IZ for initial exploratory analyses because it was one of the most recent forms available and because it had a large focal group. Large focal groups are important in estimating DIF accurately, and small samples require special DIF techniques. The results from Form IZ, which was administered to 7,405 students from California public high schools in 1999 (6,548 White students and 857 African American students), confirm the relationship observed by Freedle in the verbal test; however, this relationship was not observed by the authors in the math test.

— Verbal Test

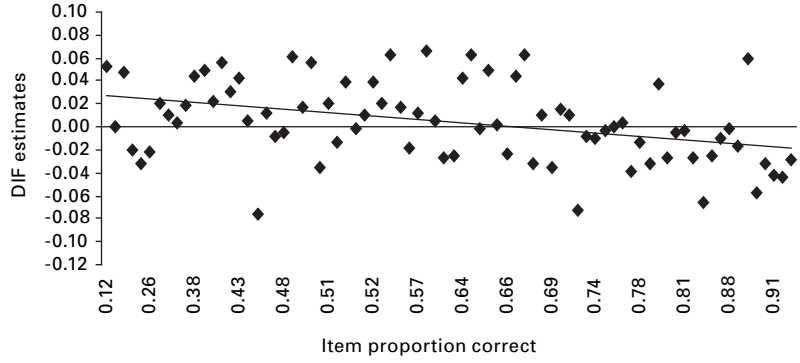
Figure 1 shows the relationship between item difficulty and DIF estimates using three different versions of the standardization approach to DIF: the standardized p-difference using denominator D_1, the standardized p-difference using denominator D_3, and the formula-score difference. In the plots, items are ordered by increasing p-value, or from hardest (fewest correct responses among all test-takers) on the left to easiest (most correct responses among all test-takers) on the right. The diagonal solid lines show the best-fit estimates using linear regression and reveal a negative relationship between the p-value and DIF estimates: easier items exhibit DIF in favor of the White group, while harder items show DIF favoring the African American group. These relationships are confirmed by the correlations between the p-values and the three DIF indices: –0.41, –0.36, and –0.42 for panels a, b, and c, respectively. These correlations are as strong as those reported by Freedle and are all statistically significant at the 0.001 alpha level.

A summary of Freedle's table 1 is presented for comparison (see table 1). Rather than p-values, Freedle used the equated delta values to measure item difficulty. It is important to note that since p-values are conceptually the opposite of item difficulty parameters like equated deltas—such that larger p-values indicate an item is relatively easier for most of the examinees in the test administration group—the positive relationship between item difficulty and DIF estimates described by Freedle corresponds to the negative relationship between p-values and DIF estimates found here. Thus, these negative correlations between p-values and DIF estimates using three versions of the standardization approach with more current test data confirm Freedle's 2003 claims.

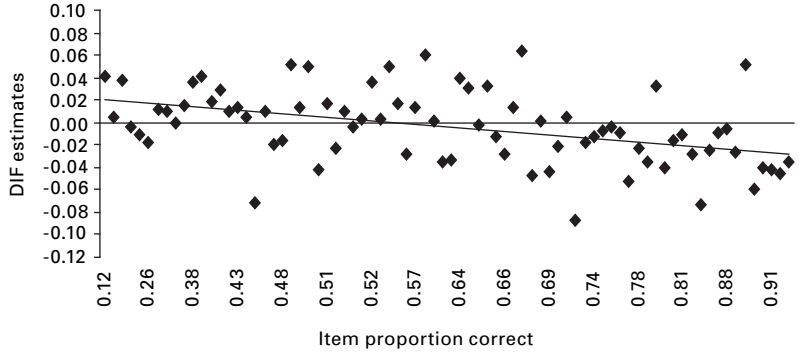
Tables 2 and 3 show the average DIF statistic for groups of items of different difficulty levels from our studies. Table 2 numerically illustrates the results shown in figure 1 by contrasting the DIF statistics for the thirty-six items in each of the hardest and easiest halves of the test as well as for the ten easiest and hardest items. Table 3 provides mean DIF for groups of ten items in increasing order of difficulty. Even if these effect sizes would be considered negligible according to Dorans and Holland (1992), it is striking that the systematic relationship between item difficulty and DIF is present just as Freedle reported. That is, the pattern of negative DIF among easier verbal items (ben-

FIGURE 1 *Relationship between item difficulty and DIF estimates using three different versions of the standardization approach in the verbal test of Form IZ 1999*

(a) Standarized p-DIF using denominator D1; items ordered by increasing p-values



(b) Standarized p-DIF using denominator D3; items ordered by increasing p-values



(c) Standarized formula-score DIF; items ordered by increasing p-values

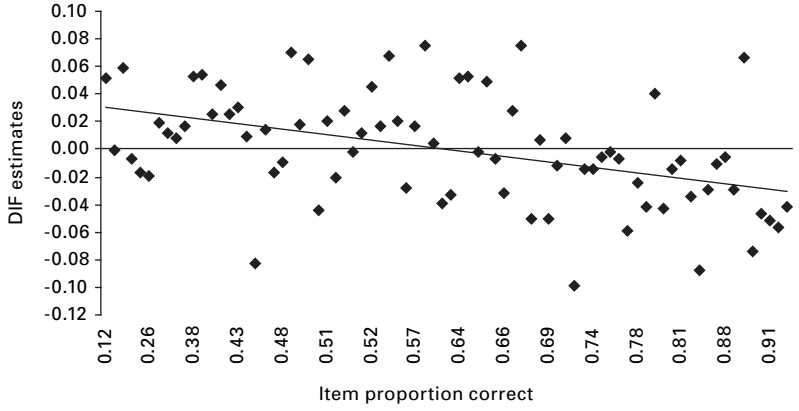


TABLE 1 *Summary of Freedle's table 1*

Item group	*Analogies*	*Antonyms*	*Sentence completion*	*Reading comprehension*
Easiest half	–0.027	–0.011	–0.010	0.005
Hardest half	0.012	0.018	0.012	0.007
Correlation between DIF and item difficulty	0.52***	0.41***	0.48***	0.08 ns

Notes: No further details about the item source were provided.

*** Significant, p < 0.001; ns = not significant (p > 0.05).

TABLE 2 *Mean DIF statistics for specific groups of items and correlations between DIF estimates and item difficulty using three different versions of the standardization approach in the verbal test of Form IZ 1999*

Item groups	*Mean p-value*	*Mean p-DIF using D_1*	*Mean p-DIF using D_3*	*Mean FS-DIF*
Easiest 10	0.8969	–0.0240	–0.0283	–0.0335
Easiest 36	0.7816	–0.0101	–0.0183	–0.0190
Hardest 36	0.4461	0.0158	0.0089	0.0159
Hardest 10	0.2591	0.0114	0.0114	0.0159
Correlation between DIF and item difficulty		–0.41***	–0.36***	–0.42***

*** Significant, p < 0.001.

efiting White students) and positive DIF among hard items (benefiting African American students) holds regardless of the standardized index employed. In addition, the DIF values observed in Form IZ are similar in size to the ones reported in Freedle's (2003) table 1, despite the use of data from more recent tests. While some of our DIF estimates match Freedle's results more closely than others, they are all comparable. These results also hold when analyzing DIF by item type (see appendix 1).

— Math Test

In contrast to the verbal test, the analyses of the math test show a weak relationship between the p-value and DIF estimates. This finding is confirmed by the correlations between the p-values and the DIF indices: –0.15, –0.09, and –0.20 when using D_1, D_3, and formula-score DIF, respectively (see table 4). None of these correlations is statistically significant at the 0.1 confidence level, and all are lower than the correlation of 0.35 obtained when comparing math item performance between White and African American students as reported by Kulick and Hu (1989) and quoted in Freedle's article.

TABLE 3 *Mean standardized indices by group of items using three different versions of the standardization approach in the verbal test of Form IZ 1999*

(a) Standardized p-DIF using D_1 as denominator

Items	*Mean p-value*	*Mean p-DIF using D_1*	*Group who benefits*	*Number of items showing 0.1>|DIF|>0.05**	*Proportion of items showing 0.1>|DIF|>0.05**	*Number of items*
1–10	0.90	–0.02	Reference	2	20%	10
11–20	0.80	–0.02	Reference	1	10%	10
21–30	0.71	–0.01	Reference	1	10%	10
31–40	0.65	0.02	Focal	2	20%	10
41–50	0.55	0.02	Focal	2	20%	10
51–60	0.50	0.01	Focal	2	20%	10
61–70	0.40	0.02	Focal	2	20%	10
71–78	0.22	0.01	Focal	1	13%	8
Total number of items				13	17%	78

(b) Standardized p-DIF using D_3 as denominator

Items	*Mean p-value*	*Mean p-DIF using D_3*	*Group who benefits*	*Number of items showing 0.1>|DIF|>0.05**	*Proportion of items showing 0.1>|DIF|>0.05**	*Number of items*
1–10	0.90	–0.02	Reference	2	20%	10
11–20	0.80	–0.03	Reference	2	20%	10
21–30	0.71	–0.02	Reference	1	10%	10
31–40	0.65	0.01	Focal	1	10%	10
41–50	0.55	0.02	Focal	1	10%	10
51–60	0.50	0.01	Focal	1	10%	10
61–70	0.40	0.01	Focal	1	10%	10
71–78	0.22	0.01	Focal	0	0%	8
Total number of items				9	12%	78

(c) Standardized FS-DIF**

Items	*Mean p-value*	*Mean FS-DIF*	*Group who benefits*	*Number of items showing medium-size DIF **	*Proportion of items showing medium-size DIF **	*Number of items*
1–10	0.90	–0.03	Reference	2	20%	10
11–20	0.80	–0.03	Reference	1	10%	10
21–30	0.71	–0.02	Reference	1	10%	10
31–40	0.65	0.01	Focal	1	10%	10
41–50	0.55	0.02	Focal	2	20%	10
51–60	0.50	0.01	Focal	2	20%	10
61–70	0.40	0.02	Focal	1	10%	10
71–78	0.22	0.01	Focal	0	0%	8
Total number of items				10	13%	78

* There are no items with DIF values larger than 0.1 or smaller than –0.1.

** Thresholds were adjusted to consider the differences in p-DIF and fs-DIF scales.

TABLE 4 *Mean DIF statistics for specific groups of items using three different versions of the standardization approach in the math test of Form IZ 1999*

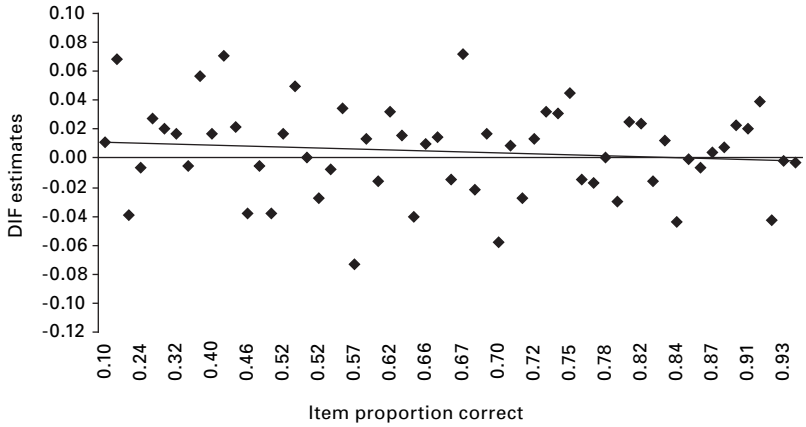
Item groups	*Mean p-value*	*Mean p-DIF using D_1*	*Mean p-DIF using D_3*	*Mean FS-DIF*
Easiest 10	0.8916	–0.0003	–0.0092	–0.0100
Easiest 30	0.7972	0.0026	–0.0049	–0.0041
Hardest 29	0.4493	0.0068	–0.0034	0.0037
Hardest 10	0.2758	0.0215	0.0049	0.0181
Correlation between DIF and item difficulty		–0.15	–0.09	–0.20

Note: Correlations are not significant at an alpha level of 0.1.

Figure 2 shows the relationship between item difficulty and DIF estimates for the math test when using the standardization approach to DIF with D_1 as the denominator. The relationship holds in other versions of the standardization approach.

The DIF values presented in table 4 are also significantly smaller than those reported by Freedle for the White/African American comparison. While the easiest half of the math form he analyzed exhibited an average standardized p-difference of –2.04 among students who reported English as their best language, the easiest half of our IZ Form showed a standardized formula-score difference of –0.0041. The hard half of the math test Freedle analyzed showed a standardized p-difference of 1.08, while the standardized formula-score

FIGURE 2 *Relationship between item difficulty and DIF estimates using the standardization approach with D1 as denominator in the math test of Form IZ 1999*



difference for Form IZ is 0.0037.[20] The other two standardization methods result in DIF estimates of similar magnitude, and none of them identifies the trend Freedle described. Only the formula-score difference shows positive DIF among hardest items and negative DIF among easiest items. The standardized p-difference, using both Freedle's denominator, D_1, and the denominator that includes all examinees, D_3, shows average DIF estimates of the same sign for the hardest twenty-nine and easiest thirty items. The weak and unstable relationship between item difficulty and DIF estimates for math items may be due to the small magnitude of the DIF statistics found in the math test.

The Freedle Phenomenon Across Test Forms and Ethnic Groups

We explored the generalizability of our findings by testing Freedle's phenomenon in four test forms (DX, QI, IZ, VD) and by ethnicity. Forms DX and QI were administered in 1994 and Forms IZ and VD in 1999. We first analyzed DIF between White and African American students and then between White and Hispanic students. All sample sizes were adequate to conduct DIF analyses (Clauser & Mazor, 1998). Table 5 includes sample size information for each ethnic group.

Because we found such low correlations between item difficulty and DIF estimates in the math test, and since most research shows only evidence of small and unsystematic DIF in the math test (Kulick & Hu, 1989; O'Neill & McPeek, 1993; Schmitt & Dorans, 1988; Zwick, 2002), we opted to focus here on only the verbal tests of the four forms.[21] Since each verbal test included 78 items, the overall analysis examined 312 verbal items.

We calculated DIF estimates using only the standardized p-difference with denominator D_3—the conservative version of the index presented by Freedle (2003)—and the standardized formula-score difference, the approach recommended by ETS researchers. Given Dorans and Zeller's (2004a) criticisms, we dropped the standardized p-difference index using denominator D_1. In addition, the high correlation we observed among our initial results from all three standardization indices suggests that findings obtained using just two of these methods are generalizable to the third.

The results presented in table 5 confirm the existence of some significant relationships between item difficulty and DIF estimates across forms. The evidence also supports the independence of the results from the form of the standardization index used. Table 5 shows that the results obtained from the standardized p-difference and the standardized formula-score difference indices are within zero to 0.02 of one another.

Although these results support Freedle's claim regarding a positive relationship between item difficulty and DIF estimates for the verbal test, they also suggest some variance in the strength of that relationship depending on the test form and focal ethnic group. For the White/African American comparison, harder items tend to benefit African American students, while easier items

TABLE 5 *Sample sizes and correlations between p-values and DIF estimates across ethnic groups and test forms in the verbal test*

		Form			
Groups	*DIF method*	*1999 IZ*	*1999 VD*	*1994 QI*	*1994 DX*
Sample sizes					
White		6,548	6,682	3,360	3,188
Hispanic		1,904	2,018	982	1,003
African American		857	929	671	709
Correlations					
White and African American	p-Value, STD P-DIF	–0.414***	–0.141	–0.317**	–0.257*
	p-Value, STD FS-DIF	–0.420***	–0.166	–0.293**	–0.240*
White and Hispanic	p-Value, STD P-DIF	–0.179	–0.101	–0.009	0.038
	p-Value, STD FS-DIF	–0.182	–0.084	0.020	0.038

*** Statistically significant ($p < 0.001$).
** Statistically significant ($p < 0.01$).
* Statistically significant ($p < 0.05$).

tend to benefit White students, regardless of the DIF method used. Estimates ranged between 0.141 and 0.420 and were close in magnitude to the correlations reported by Freedle (2003), although the results for Form VD were not statistically significant. By contrast, the Freedle phenomenon was not evident in the comparison between Hispanic and White students. None of the correlations from the White/Hispanic item comparison reached the magnitude reported by Freedle (see table 6),[22] and none was statistically significant. Furthermore, there is an almost null relationship between item difficulty and DIF estimates for the White/Hispanic item comparison in the two forms administered in 1994.

The extent to which items exhibited serious and medium-sized DIF was also analyzed across forms and ethnic groups with the goal of shedding some light on the correlation pattern observed in table 5. Not surprisingly, higher correlations are observed in test forms with more items exhibiting serious or medium-sized DIF (items with DIF estimates above |0.5|).

Overall, the test forms present a number of items exhibiting serious or medium-sized DIF (see table 7). This is interesting, especially considering that we expected to find very few items in any of these categories since the test forms had undergone the DIF screening procedure. These results suggest that although the procedure may help to minimize the presence of items with large DIF, it has not succeeded at entirely eliminating DIF from the SAT—particularly for African American test-takers.

TABLE 6 *Summary of Freedle's table 5 showing correlations between DIF estimates and equated delta measures of item difficulty for analogies and sentence completion items*

Ethnic groups	*Analogy correlation*	*Sentence completion correlation*
African American	0.496***	0.275**
Hispanic American	0.555**	0.263*

*** Significant, $p < 0.001$.
** Significant, $p < 0.01$.
* Significant, $p < 0.05$.

TABLE 7 *Classification of DIF items*

		SAT form							
		1999 IZ		*1999 VD*		*1994 QI*		*1994 DX*	
Ethnic group comparison	*STD DIF index*	*serious**	*TBI***	*serious**	*TBI***	*serious**	*TBI***	*serious**	*TBI***
White and African American	STD p-DIF	0	9	0	10	0	11	0	13
	STD FS-DIF	0	10	0	10	0	14	0	12
White and Hispanic	STD p-DIF	1	3	0	7	1	3	0	4
	STD FS-DIF	1	3	0	4	1	3	0	4

* Serious: In the standardization method, serious DIF is defined as standardized indexes over 0.1 or below –0.1. Thresholds were adjusted to consider the differences in P-DIF and FS-DIF scales.

** TBI: To-be-inspected items in the standardization method are defined as those with standardized indexes between 0.1 and 0.05 and –0.05 and –0.1. Thresholds were adjusted to consider the differences in p-DIF and FS-DIF scales.

Discussion

The relationship we found between item difficulty and DIF estimates for the White/African American comparison of verbal items falls between the correlations reported by Dorans and Zeller (2004a) and the correlations found earlier by ETS researchers (Burton & Burton, 1993; Freedle, 2003; Kulick & Hu, 1989). Although, on average, the relationship is weaker than the 0.50 reported by Freedle, three of the four forms exhibit correlations above the 0.19 reported by Dorans and Zeller between equated deltas and the standardization formula-score difference index. This relationship holds consistently across different versions of the standardization approach to DIF and also in more current forms than those examined in Freedle's paper. We found the highest correlation (0.4) between DIF estimates and p-values in a 1999 form (IZ), one of the most current test forms analyzed. The findings from the verbal tests suggest that the pretest screening, to which all SAT items have been submitted since the early 1980s, has not reduced the degree of correlation

between DIF and difficulty among verbal items as much as Dorans and Zeller have claimed.

We did not find evidence, however, that this is a widespread phenomenon occurring among all kinds of multiple choice questions and affecting all groups other than the "White English speaker majority" (Freedle, 2003). The correlation between item difficulty and DIF estimates is weak to nonsignificant for the White/African American comparison of math items and for the White/Hispanic comparison of both verbal and math items. Hence, ETS's pretest screening for math items seems to be effective.

These results likely hold for current test forms, because the changes made to the SAT verbal and math sections in 2005 were minor and revisions centered on the introduction of a writing test (Mattern et al., 2008). Further, our item-type analysis (presented in appendix 1) shows that the general phenomenon holds in the sentence completion and reading comprehension sections, which were maintained in the test beyond 2005. Dropping the analogies section may potentially help reduce the phenomenon observed in the verbal test, but it is unlikely to eliminate it. We also accounted for guessing in our analysis, as advised by critics, but this did not change our main conclusion: we continue to see a relationship between DIF and item difficulty in the verbal test.

Although our findings limit the phenomenon to the verbal test and the African American subgroup, these findings are important because they show that the SAT, a high-stakes test with significant consequences for the educational opportunities available to young people in the United States, favors one ethnic group over another. Neither the specifics of the method used to study differential item functioning nor the date of the test analyzed invalidate Freedle's claims that the SAT treats African American minorities unfairly.

The confirmation of unfair test results throws into question the validity of the test and, consequently, all decisions based on its results. All admissions decisions based exclusively or predominantly on SAT performance—and therefore access to higher education institutions and subsequent job placement and professional success—appear to be biased against the African American minority group and could be exposed to legal challenge.

Despite these critical findings, our analyses do not enable us to elaborate on the causes behind these results. We do not know if Freedle's (2003) hypothesis about cultural and linguistic differences, which was based on the work of Diaz-Guerrero and Szalay (1991), is driving the results or whether the systematic relationship between DIF and item difficulty is explained by some other cause. Our methodology also does not allow us to explain why there is White/African American DIF and not White/Hispanic DIF.

The confirmation of this systematic phenomenon in the verbal test suggests areas where further inquiry is needed. Subsequent research should aim to understand more completely the basis of the observed relationship between item difficulty and DIF as well as possible measures that might be taken to minimize its effects across population subgroups. As performance on the SAT

has dramatic consequences for students in the United States, the validity and fairness of the SAT should be held to a very high standard.

ETS, the organization responsible for developing the test, should take additional action to investigate the causes behind the findings presented here. Some possibilities include conducting think-alouds among students from the White and African American subgroups to study their cognitive processes when reading and responding to the test and to examine, as Freedle (2003) suggested, whether hard verbal items are a more valid assessment of African American students' knowledge and skills than easier verbal items. The information provided by examinees in such in-depth interviews, which could be conducted while they complete the test or right after, would help researchers and test developers understand the role of language and content in this phenomenon. That information would in turn help test developers and item writers devise items with content and language that is neutral to members of different ethnic subgroups. Another possibility is to explore the causes behind this issue using recently developed quantitative methodologies, such as exploratory item response models that allow researchers to relate item functioning to item's and student's observed characteristics. While this issue is still being addressed, colleges and universities should be strongly advised not to consider verbal SAT scores in isolation when making admissions decisions.

Conclusion

The findings from this research confirm the relationship between item difficulty and DIF estimates reported by Freedle for the African American/White comparison of verbal items. The relationship is found in more current test forms than those analyzed by Freedle and persists after addressing the methodological issues posed by ETS researchers. However, this study did not find evidence to suggest that this issue applies to Hispanic students, nor did it find evidence to suggest that the issue applies to questions other than verbal items.

Our confirmation of Freedle's (2003) findings using different methodology, data from more current tests, and controls for the role of guessing, is important because Freedle's claims of biased SAT results for minority students were strongly questioned on these grounds. As independent researchers, we have objectively addressed the criticisms of Freedle and found that his findings still hold. Settling this argument is critical so that we can move forward and investigate the causes behind differential item functioning between African American and White SAT test-takers. Since 2003, the questions Freedle raised and the debate he initiated have been set aside for methodological reasons. Tragically, the dismissal of his work has stopped involved and concerned parties from asking and discussing substantive, challenging questions about fairness in access to higher education. Now that we have confirmed that the relationship between item difficulty and DIF exists, and that it is not just an artifact of methodology, we hope we have opened the space necessary to advance our

understanding of what lies beneath this relationship and what should be done (and not done) with test results that are invalid for some subgroups of the population. Although not generalizable to all ethnic subgroups and to all item types, these findings are certainly strong enough to question the validity of SAT verbal scores for African American examinees and consequently admission decisions based exclusively or predominantly on those scores.

Ensuring that test items are fair for all subgroups is the professional and ethical duty of all test developers. The SAT continues to be one of the most influential tests in the United States. The fairness of its results should be of utmost importance not only to students and those directly related to test development but also to anyone concerned with promoting fair access to higher education for different ethnic groups.

Notes

1. When we study DIF based on information coming from the same test (items) under analysis, we talk about an *internal-reference criterion*. When an external variable, different from test performance, is used to study test or item functioning in different subgroups of the population, we talk about *external-reference criterion* or *differential prediction*.
2. *Ability* refers to the amount of a variable that an examinee has. This variable is often latent or unobserved and refers to the construct of interest. In this case, *ability* refers to the student's total score and is an observed variable.
3. While proportion-correct scores consider only correct responses to compute a total score, formula scoring adjusts the total score by the probability that a student may get a correct response as result of guessing. We provide more detail on the definition of both these concepts in the methodology section.
4. Students complete this questionnaire when they register for the SAT.
5. These are items that explicitly include references to minorities (e.g., literature, activities, celebrities). In these items, the focal group performs relatively better.
6. Verbal items are generally classified into four areas: aesthetic/philosophy, practical affairs, science, and human relationships.
7. The equated delta is a measure of item difficulty expressed in the delta scale used extensively at ETS. The delta metric is obtained from the percentage correct through an inverse normal transformation and is scaled to have a mean of thirteen and a standard deviation of four. On this scale, delta increases for more difficult items and decreases for easier items. Since delta statistics are population dependent, equating is used to place them on a common scale across test forms (Kulick & Hu, 1989).
8. Freedle and Kostin (1991) explain that "in each set of ten analogies, the first item is typically the easiest and tenth item is the hardest" (p. 3).
9. Nonspeeded items are items to which examinees respond presumably with no time restriction. Some studies have considered items in the middle of a section, which examinees may answer under less time pressure, as nonspeeded.
10. Freedle's (2003) scoring at the item level excluded examinees who did not respond to the question, and his averages are therefore based on an indeterminate group.
11. The particular test forms he used were administered prior to the time that the procedure for screening out items exhibiting DIF was put in place.
12. Bridgeman and Burton (2005) deemed the R-SAT score invalid because it does not correlate with other scales that should be measuring the same construct, such as high school grade-point average.
13. Refer to Santelices (2007) for an alternative approach using item response modeling.

14. The standardization approach also permits the analysis of distractor, not-reached, and omitted responses (Schmitt & Bleistein, 1987).
15. Uniform or consistent DIF occurs when the item characteristic curves for two groups are different and do not cross. There is a relative advantage for one group over the entire ability range. Nonuniform or inconsistent DIF occurs if the item characteristic curves for two groups are different but cross at some point on the θ scale. Therefore, nonuniform DIF for and against a given group balance or cancel each other to some degree (Camilli & Shepard, 1994).
16. Guidelines about the thresholds for large, medium, and negligible DIF effect size should be adjusted when using the standardized formula-score DIF index in order to take into account the difference in scales between the p-difference index and the fs-difference index. The adjusted thresholds were used in the analyses presented in the results section of this paper.
17. The two forms from each year contain the same items but in different order. The four forms were chosen by the College Board to reflect the usual variation between the versions of the SAT normally used in a given year and between years.
18. Although Freedle (2003) made the distinction between examinees who spoke English as their "preferred language" and those who didn't, previous research distinguished between examinees who spoke English as their "best language" and those who did not. We decided to implement this latter tradition.
19. President Atkinson's interest in the SAT resulted in, among other things, the modifications made to the test in 2005 (Atkinson & Pelfrey, 2004).
20. Although Freedle (2003) presents the pattern of DIF estimates for easy and hard math items, he does not present the correlation between DIF estimates and item difficulty, presumably because he no longer had access to the data.
21. We also conducted the analyses on the math test from Forms DX, QI, IZ, and VD for both the White/African American and the White/Hispanic comparisons. The results show that the phenomenon described by Freedle (2003) is evident in the 1994 forms, especially for the White/African American comparison, but is not present in more current math tests. No results for the 1999 forms are statistically or practically significant ($p < 0.05$). See appendix 2 for details.
22. Freedle (2003) did not report the overall test correlation between item difficulty and DIF estimates by ethnic group. He only reported correlations by item type across ethnic groups. In addition, it is important to remember that rather than p-values, Freedle used the equated delta values to measure item difficulty. Therefore, the relationship he described is consistent with a positive correlation between DIF and item difficulty and a positive relationship between p-values and DIF.

References

AERA, APA, & NCME. (1999). *Standards for educational and psychological testing*. Washington, DC: AERA.

Atkinson, R., & Pelfrey, P. (2004). *Rethinking admissions: US public universities in the post–affirmative action age*. Retrieved March 25, 2008, from http://cshe.berkeley.edu/publications/publications.php?s=1

Brennan, R. (Ed.). (2006). *Educational measurement*. Westport, CT: Praeger.

Bridgeman, B., & Burton, N. (2005, April). *Does scoring only the hard questions on the SAT make it fairer?* Paper presented at the annual meeting of the American Educational Research Association, Montreal.

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. 321–336). Hillsdale, NJ: Erlbaum.

Camilli, G., & Shepard, L. (1994). *Methods for identifying biased test items* (Vol. 4). London: Sage.

Carlton, S., & Harris, A. (1992). *Characteristics associated with differential item functioning on the Scholastic Aptitude Test: Gender and majority/minority group comparisons* (No. RR-92-64). Princeton, NJ: Educational Testing Service.

Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items. In G. Engelhard (Ed.), *ITEMS. Instructional Topics in Educational Measurement.* National Council on Measurement in Education.

College Board. (2008). College-bound senior press-release. Retrieved from http://preofessionals.collegeboard.com/profdownload/CBS-2008-press-release-FINAL.pdf

Diaz-Guerrero, R., & Szalay, L. B. (1991). *Understanding Mexicans and Americans: Cultural perspectives in conflict.* New York: Plenum Press.

Dorans, N. (2004). Freedle's table 2: Fact or fiction. *Harvard Educational Review, 74*(1), 62–79.

Dorans, N., & Holland, P. (1992). *DIF detection and description: Mantel-Haenszel and standardization* (No. RR-92-10). Princeton, NJ: Educational Testing Service.

Dorans, N., & Kulick, E. (1983). *Assessing unexpected differential item performance of female candidates on SAT and TSWE forms administered in December 1977: An application of the standardization approach* (No. RR-83-9). Princeton, NJ: Educational Testing Service.

Dorans, N., & 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*(4), 355–368.

Dorans, N., & Lawrence, I. (1987). *The internal construct validity of the SAT* (No. RR-87-35). Princeton, NJ: Educational Testing Service.

Dorans, N., & Zeller, K. (2004a). *Examining Freedle's claims and his proposed solution: Dated data, inappropriate measurement, and incorrect and unfair scoring* (No. RR-04-26). Princeton, NJ: Educational Testing Service.

Dorans, N., & Zeller, K. (2004b). *Using score equity assessment to evaluate the equatability of the hardest half of a test to the total test* (No. RR-04-43). Princeton, NJ: Educational Testing Service.

Frary, R. (1988). Formula scoring of multiple-choice tests (correction for guessing). In B. S. Plake (Ed.), *ITEMS. Instructional Topics in Educational Measurement.* National Council on Measurement in Education.

Freedle, R. (2003). Correcting the SAT's ethnic and social-class bias: A method for reestimating SAT scores. *Harvard Educational Review, 73*(1), 1–44.

Freedle, R., & Kostin, I. (1988). *Relationship between item characteristics and an index of differential item functioning for four GRE verbal item types.* Princeton, NJ: Educational Testing Service.

Freedle, R. & Kostin, I. (1991). *Semantic and strutural factors affecting the performance of matched black and white Scholastic Aptitude Test.* (No. RR 91-25). Princeton, NJ: Educational Testing Service.

Freedle, R., Kostin, I., & Schwartz, L. M. (1987). *A comparison of strategies used by black and white students in solving SAT verbal analogies using a thinking aloud method and a matched percentage-correct design* (No. RR 87-48). Princeton, NJ: Educational Testing Service.

Holland, P., & Thayer, D. T. (1988). Differential item performance and Mantel-Haenszel. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 129–145). Hillsdale, NJ: Erlbaum.

Joint Committee on Testing Practices. (2004). *Code of fair testing practices in education.* Washington, DC: Joint Committee on Testing Practices.

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., Curley, W. E., & McHale, F. (1988). *Differential item functioning for males and females on SAT-verbal reading subscores items.* (No. RR-88-10). Princeton, NJ: Educational Testing Service.

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

Mattern, K., Patterson, B., Shaw, E., Kobrin, J., & Barbuti, S. (2008). *Differential validity and prediction of the SAT*. New York: College Board.

O'Neill, K., & McPeek, W. M. (1993). Item and test characteristics that are associated with differential item functioning. In P. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 255–276). Hillside, NJ: Erlbaum.

Ramsey, P. (1993). Sensitivity review: The ETS experience as a case study. In P. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 367–388). Hillsdale, NJ: Erlbaum.

Santelices, M. V. (2007). Differential item functioning in the SAT I: Reasoning Test. Unpublished doctoral dissertation. University of California Berkeley, Berkeley.

Schmitt, A., & Bleistein, C. (1987). *Factors affecting differential item functioning for black examinees on Scholastic Aptitude Test analogy items.* (No. RR-87-23). Princeton, NJ: Educational Testing Service.

Schmitt, A., & Dorans, N. (1988). *Differential item functioning for minority examinees on the SAT* (No. RR-88-32). Princeton, NJ: Educational Testing Service.

Zieky, M. (1993). Practical questions in the use of DIF statistics in test development. In P. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 337–348). Hillside, NJ: Erlbaum.

Zwick, R. (2002). *Fair game? The use of standardized admissions tests in higher education*. New York: Routledge Falmer.

Zwick, R., Donoghue, J. R., & Grima, A. (1993). Assessment of differential item functioning for performance tasks. *Journal of Educational Measurement, 30*(3), 233–251.

We thank Saul Geiser for helpful discussions, the College Board for providing the data, and University of California All Campus Consortium on Research for Diversity (UC ACCORD) for granting funding through a dissertation fellowship.

Appendix 1

The general results observed in the verbal section of the SAT Form IZ are also observed when we analyze verbal items by type (see table A1). The relationship between item difficulty and DIF is present in the reading comprehension and sentence completion items regardless of the DIF method used. It is important to note that our results by item type are somewhat different from the ones obtained by Freedle (2003). While Freedle observed a relationship between item difficulty and DIF estimates in three of the four item types he studied (analogies, antonyms, and sentence completion), we observe the phenomenon in only two of the three item types included in the SAT in 1999 (sentence completion and reading comprehension). Where Freedle found a very strong pattern among analogies, we find none. In addition, Freedle found no relationship between item difficulty and DIF estimates among reading comprehension items, but we do observe results supporting this relationship. The sentence completion items also show the relationship Freedle reported.

TABLE A1 *Mean standardized indices by item type using three different versions of the standardization approach in the verbal test of Form IZ 1999*

(a) Standardized p-index using D_1 as denominator

Item type		*Mean p-value*	*Mean p-DIF using D_1*	*Group who benefits*	*Number of items with 0.1>|DIF|>0.05**	*Proportion of items with 0.1>|DIF|>0.05**	*Number of items*
Analogies (13)	Easier 5	0.87	–0.01	Reference	2	40%	5
	Harder 5	0.37	–0.02	Reference	1	20%	5
Reading comprehension (40)	Easier 5	0.84	–0.01	Reference	0	0%	5
	Middle 5	0.60	0.03	Focal	2	40%	5
	Harder 5	0.35	0.04	Focal	1	20%	5
Sentence completion (25)	Easier 5	0.90	–0.04	Reference	1	20%	5
	Middle 5	0.57	0.01	Focal	1	20%	5
	Harder 5	0.21	0.00	–	1	20%	5
Total number of items					9	23%	40

(b) Standardized p-index using D3 as denominator

Item type		*Mean p-value*	*Mean p-DIF using D_3*	*Group who benefits*	*Number of items showing 0.1>|DIF|>0.05**	*Proportion of items with 0.1>|DIF|>0.05**	*Number of items*
Analogies (13)	Easier 5	0.87	–0.02	Reference	2	40%	5
	Harder 5	0.37	–0.02	Reference	1	20%	5
Reading comprehension (40)	Easier 5	0.84	–0.02	Reference	0	0%	5
	Middle 5	0.60	0.02	Focal	1	20%	5
	Harder 5	0.35	0.02	Focal	0	0%	5
Sentence completion (25)	Easier 5	0.90	–0.04	Reference	1	20%	5
	Middle 5	0.57	0.00	–	0	0%	5
	Harder 5	0.21	0.01	Focal	0	0%	5
Total number of items					5	13%	40

(c) Standardized FS-index**

Item type		*Mean p-value*	*Mean FS-DIF*	*Group who benefits*	*Number of items showing medium size DIF**	*Proportion of items medium size DIF **	*Number of items*
Analogies (13)	Easier 5	0.87	–0.02	Reference	2	40%	5
	Harder 5	0.37	–0.02	Reference	1	20%	5
Reading comprehension (40)	Easier 5	0.84	–0.02	Reference	0	0%	5
	Middle 5	0.60	0.03	Focal	1	20%	5
	Harder 5	0.35	0.04	Focal	0	0%	5
Sentence completion (25)	Easier 5	0.90	–0.05	Reference	1	20%	5
	Middle 5	0.57	0.00	–	1	20%	5
	Harder 5	0.21	0.01	Focal	0	0%	5
Total number of items					6	15%	40

*There are no items with DIF values larger than 0.1 or smaller than –0.1.

**Thresholds were adjusted to consider the differences in p-DIF and FS-DIF scales.

Appendix 2

TABLE A2 *Sample size and correlations between DIF estimates and p-values using two different versions of the standardization approach in the math test across ethnic groups and test forms*

		Form			
Group	*DIF method*	*1999 IZ*	*1999 VD*	*1994 QI*	*1994 DX*
Sample size					
White		6,548	6,682	3,360	3,188
Hispanic		1,904	2,018	982	1,003
African American		857	929	671	709
Correlations		857	929	671	709
White and African American	p-value, STD P-DIF	–0.09	–0.08	–0.33*	–0.28*
	p-value, STD FS-DIF	–0.20	–0.18	–0.36**	–0.28*
White and Hispanic	p-value, STD P-DIF	0.01	0.00	–0.17	–0.05
	p-value, STD FS-DIF	–0.11	–0.17	–0.26*	–0.01

*** Statistically significant ($p < 0.001$).

** Statistically significant ($p < 0.01$).

* Statistically significant ($p < 0.05$).