CollegeBoard
connect to college success™

College Board Response to *Harvard Educational Review* Article by Freedle

Wayne Camara and Viji Sathy
April 2004

Roy O. Freedle's recent article in the *Harvard Educational Review*, entitled "Correcting the SAT's Ethnic and Social-Class Bias: A Method for Reestimating SAT® Scores," is based on small differences between white students' responses and the responses of students from other ethnic groups to test items that were discussed by a number of researchers (Angoff & Ford, 1973; Burton & Burton, 1993; Kulick & Hu, 1989; Schmitt & Bleistein, 1987), as well as by Freedle and his associates (Freedle & Kostin, 1988, 1990, 1997). Freedle claims that his revised measure will create a valid measure that reduces group differences significantly. Although any study that purports to reduce group differences must be looked at seriously, Freedle's study is so flawed that its conclusions are misleading.

There are myriad technical problems with the report, including misuse of regression and differential item functioning (DIF), and even a misunderstanding of how scores on the SAT are calculated. But one need not be a psychometrician to understand the fundamental problem with the study. The reduction in group differences is not the result of more sensitive or appropriate measurement but, rather, it is because the proposed measure relies mostly on students guessing the answers to test questions.

To probe a little deeper, let us examine more closely Freedle's argument around DIF. Researchers have found that, on average, African American, Hispanic, and Asian American students tend to choose the correct response on easy test questions slightly less often than white students with an equal total test score. In contrast, they choose the correct response on difficult test questions slightly more often than white students with an equal total test score. Noting that this phenomenon occurs with SAT vocabulary questions but not with critical reading questions, Freedle suggests that the College Board should dispense with SAT critical reading questions, as well as the easier half of all vocabulary questions, to improve the scores of ethnic minority test-takers.

The suggestion that critical reading be dropped or de-emphasized on the SAT, given its importance for success in college, would not be educationally or psychometrically sound even if it were based on a credible analysis. Not only is critical reading a major component of the verbal reasoning construct but the SAT verbal section also has been shown to predict success in college for all students, including ethnic and racial minority students, men and women, students with disabilities, and native and non-native English speakers (Bridgeman, Jenkins, & Ervin, 2000; Ragosta, Braun, & Kaplan, 1991; Ramist, 1984; Ramist, Lewis, & McCamley-Jenkins, 1994). Freedle himself notes that the critical reading items lack what he calls "the familiar pattern of bias."

Let us look briefly at the data for the so-called SAT-R Section that Freedle recommends. On the difficult items that are included in the SAT-R, African American candidates receive an average score of 22 percent out of a perfect score of 100 percent. Since there are five answer options for each question, 22 percent is only slightly above what would be expected from random guessing, namely 20 percent. White candidates do somewhat better, achieving an average score of 31 percent. The results indicate that this test is too hard for either group and would be a frustrating experience for most students. There are simply too many questions that are geared to those with a much higher level of knowledge and skill than is required of college freshmen. Extending Freedle's argument, we could substantially reduce all group differences if the test were made significantly more difficult so that all examinees would have to guess the answers to nearly all of the questions. We could then predict that each subgroup would have an average of 20 percent of their answers correct, based on chance.

Freedle proposes a post-hoc explanation for these smaller differences on hard questions by appropriating the work of Diaz-Guerrero & Szalay (1991) to suggest that the reason for the reduced differences on his measure is the result of the ways that individuals understand words and concepts. Freedle overgeneralizes from an analysis that finds cultural differences in words that have particular social and political meaning to posit widespread cultural differences in all languages. This explanation is even less plausible in explaining similar differences observed for mathematics items, since the difficulty of math items is a function of the mathematical content—the vocabulary employed in mathematics items tends to be simple at all difficulty levels.

In brief, Freedle's suggestions boil down to capitalizing on chance performance. This kind of performance may represent either random guesses, or unconnected bits of knowledge that are not sufficiently organized to be of any use in college studies. With random guessing, some students *will* receive large windfalls, but some will have equally bad luck, and none will earn reliable scores. If students who responded randomly took the test again, their performance would be inconsistent with their earlier performance and would also be unrelated to performance in college.

In our view, Freedle has presented no argument that would justify removing or de-emphasizing a critical college success skill, such as reading, from the SAT. Nor has he presented a credible case for constructing tests that are very difficult for the intended population of examinees.

References

Angoff, W. H., & Ford, S. F. (1973). Item-race interaction on a test of scholastic aptitude. *Journal of Educational Measurement*, *10*, 95–106.

Bridgeman, B., Jenkins, L., & Ervin, N. (2000). *Predictions of freshman grade point average from the revised and recentered SAT I: reasoning test.* College Board Report No. 2000-01. New York: College Board.

Burton, E. B., & 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*. Hillsdale, NJ: Erlbaum.

Diaz-Guerrero, R., & Szalay, L. B. (1991). *Understanding Mexicans and Americans: Cultural language of maternal control*. London: Routledge & Kegan Paul.

Freedle, R., & Kostin, I. (1988). *Relationship between item characteristics and an index of Differential Item Functioning (DIF) for the four GRE verbal item types*. ETS RR-88-29. Princeton, NJ: Educational Testing Service.

Freedle, R., & Kostin, I. (1990). Item difficulty of four verbal item types and an index of Differential Item Functioning for black and white examinees. *Journal of Educational Measurement*, *27*, 329–343.

Freedle, R., & Kostin, I. (1997). Predicting black and white differential item functioning in verbal analogy performance. *Intelligence*, *24*, 417–444.

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

Ragosta, M., Braun, H., & Kaplan, B. (1991). *Performance and persistence: A validity study of the SAT for students with disabilities*. College Board Report No. 91-3. New York: College Board.

Ramist, L. (1984). Validity of the ATP tests. In T. Donlon (ed.), *The College Board technical handbook for the Scholastic Aptitude Test and Achievement Tests* (pp. 141–170). New York: College Board.

Ramist, L., Lewis, C., & McCamley-Jenkins, L. (1994). *Student group differences in predicting college grades: Sex, language, and ethnic group.* College Board Report No. 93-1. New York: College Board.

Schmitt, A. P., & Bleistein, C. A. (1987). *Factors affecting differential item functioning for black examinees on Scholastic Aptitude Test analogy items*. ETS RR-87-23. Princeton, NJ: Educational Testing Service.