Responding to Claims of Misrepresentation

MARIA VERONICA SANTELICES
Pontificia Universidad Católica de Chile

MARK WILSON
University of California, Berkeley

In our paper "Unfair Treatment? The Case of Freedle, the SAT, and the Standardization Approach to Differential Item Functioning" (Santelices & Wilson, 2010), we studied claims of differential effects of the SAT on Latinos and African Americans through the methodology of differential item functioning (DIF). Previous research (Freedle, 2003) identified a systematic relationship between item difficulty and DIF results in the SAT: harder items tended to systematically benefit minority students while easier items benefited white students. The systematic phenomenon was explained by a cultural and linguistic hypothesis. Our article investigated the relationship between item difficulty and DIF by replicating and expanding on this highly controversial previous research. Our analysis addressed criticisms (Dorans, 2004; Dorans & Zeller, 2004) against Freedle's 2003 article by analyzing data from more recent test forms, considering the effect of no responses in the DIF methodology, and considering the possibility of guessing in the scoring used. This was done using the standardization approach to DIF. The results confirmed the relationship in the Verbal test for the White/African American comparison. There is no evidence in the study, however, that this phenomenon occurred in Math test items, nor was it observed in the White/Latino comparison.

We begin this brief rejoinder to the Dorans and Freedle responses by agreeing with Dorans: the main focus of our article is the relationship between item difficulty and DIF. This is the main statistical phenomenon described in Freedle's 2003 article: "Whites tend to score better on *easy* items and African Americans on *hard* items" (p. 3). Our article did not deal with Freedle's second contribution, which was to introduce the Revised-SAT (R-score). Thus, Dorans's criticism that we failed to address Freedle's concern "about the miscalculation of percent correct for the hard-half test and the efficacy of the R-score" (p. 409) is not relevant to our argument. We were not addressing that issue.

We believe that the connection between DIF/difficulty correlations and fairness is a matter worth investigating, despite claims to the contrary by Dorans in his response to our article (p. 409). Our judgment that it is important is borne out by the existence of a literature on the topic, including earlier work by Dorans himself (Dorans & Zeller, 2004; Kulick & Hu, 1989; Rogers & Kulick, 1986; Shepard, Camilli, & Williams, 1984). For example, in 2004 Dorans and Zeller wrote "there appears to be a DIF/difficulty relationship that merits some investigation" (p. 27). Furthermore, two subsequent papers have referred to Freedle's initial publication by describing and addressing exactly this statistical relationship: Wainer and Skorupski (2005) and Scherbaum and Goldstein (2008). We agree with them that it is an important and interesting scientific topic. Even if there is instability in these correlations as Dorans claims in his response to our article (p. 408), that is not sufficient to eliminate concerns about unfairness, as the evidence for instability itself relies on the correlation being observed as non-zero in some contexts. The main conclusion we drew in our article was that further research is needed on this connection and, specifically, that we see a need to investigate the impact of this phenomenon on fairness issues. This new research should not be limited to the analysis of DIF and item content as suggested by Freedle in his response to our article (pp. 395–396); it also should target the relationship between DIF and item difficulty using quantitative analyses and modeling techniques and incorporating content analyses where appropriate. It is the relationship between item difficulty and DIF that is our major interest, even if DIF magnitudes are small or medium.

Dorans also claims that we have made misrepresentations, substituting "*considered serious* in place of *more unusual*" (p. 407). In fact, we used the words *considered serious* as *our* words to describe the gravity of the findings when the absolute value of the standardization statistic is over 0.1. Here is what some researchers have said about this particular range of DIF effect sizes: "[they] are more unusual and should be examined very carefully" (Dorans & Holland, 1993, p. 50); "[they] require careful examination that sometimes leads to the conclusion that the item is biased" (Schmitt & Dorans, 1988, p. 8); "…|DSTD|>=.10 flags relatively few items, most of which are problematic" (Dorans & Kulick, 1986, p. 361). We think the word *serious* accurately describes the situation to which these researchers refer.

Subsequently, Dorans claims we have misrepresented the number of SAT forms we used (p. 408). In fact, the four SAT forms we analyzed were given to us already packaged and named (as forms IZ, VD, QI, and DX) by the College Board. These are comprised of two different item sets and two different orderings of each item set. The results coming from two forms with items in different order will have different response patterns and different statistical features, although variation could well be less than between two forms with different sets of items (Jansen & Kebede, 2009; Kingston & Dorans, 1982; Schmitt & Bleistein, 1987; Schweizer, 2009). Analyzing aggregated responses by item, regardless of item position, would have gone against the findings of the literature on this subject (Jansen & Kebede, 2009; Schweizer, 2009). In addition, and as originally stated in our article, the sample sizes used in the study are within the range suggested by Clauser and Mazor (1998).

While predictive validity has merit in evaluating the performance of tests in general, and of the SAT specifically, we emphasize that DIF is also appropriate in examining bias: studying DIF is one of the methodologies that the Standards for Educational and Psychological Testing use to define bias and is part of "a sound testing practice" (AERA, APA, & NCME, 1999, p. 79). Dorans argues, however (wrongly, we believe), that DIF is exclusively a test construction tool (p. 411).

Dorans also accuses us of misunderstanding previous criticisms (p. 408), of designing a study to demonstrate the obvious (p. 408), and of showing confirmation bias (p. 410). We think these are unsupported claims. Note the following statement by Dorans and Zeller (2004): "DIF screening seems to have successfully reduced the degree of correlation between DIF and [item] difficulty that served as the impetus for Freedle`s provocative claims" (p. 26). This is exactly the sort of claim our study set out to answer. The correlation between item difficulty and DIF was at the heart of Freedle's argument. Our article confirms Freedle's (2003) findings of a systematic relationship between item difficulty and DIF using the same methodology but implementing modifications that researchers suggested back in 2004 (Dorans, 2004; Dorans & Zeller, 2004).

According to our results, however, the relationship holds in a more circumscribed situation than the one described by Freedle (2003): only between Whites and African Americans and only in the Verbal test. We do not find evidence to support that the correlation described by Freedle (2003) is present in the Math test of these forms, nor is it observed in the Hispanic/White comparison. We think the pattern cannot at this point be generalized to all standardized tests and all ethnic minority groups. In addition, and similar to Freedle's original findings, the DIF magnitudes observed are small or medium.

Our aim has been to contribute to the discussion of the relationship between item difficulty and DIF, in part by addressing several of the methodological considerations raised seven years ago. We know now that the relationship still holds, at least in some contexts, and we think the appropriate thing to do is move on to investigating its causes and potential impact on total test scores and real-life decisions made based on those scores.

References

American Educational Research Association [AERA], American Psychological Association [APA] & National Council on Measurement in Education [NCME] (1999). *Standards for educational and psychological testing.* Washington, D.C.: AERA.

Clauser, B. E., & Mazor, K. M. (1998). *Using statistical procedures to identify differentially functioning test items.* Retrieved March 10th, 2006, from www.ncme.org.

Dorans, N. (2004). Freedle's Table 2: Fact or fiction. *Harvard Educational Review, 74*(1), 62–72.

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

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., & Zeller, K. (2004). *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.

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

Jansen, R., & Kebede, M. (2009). *Modeling item order effects within a DIF framework.* Paper presented at the Annual Meeting of the Psychometric Society, Cambridge, UK.

Kingston, N. M. & Dorans, N. (1982). *The effect of an item position within a test on item responding behavior: An analysis based on item response theory* (No. RR-82-22). Princeton, NJ: Educational Testing Service.

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

Rogers, H. J., & Kulick, E. (1986). *An investigation of unexpected differences in item performance between blacks and whites taking the SAT.* Paper presented at the National Council on Measurement in Education, San Francisco.

Santelices, M. V., & Wilson, M. (2010). Unfair treatment? The case of Freedle, the SAT, and the standardization approach to differential item functioning. *Harvard Educational Review, 80*(1), 106–133.

Scherbaum, C., & Goldstein, H. (2008). Examining the relationship between race-based differential item functioning and item difficulty. *Educational and Psychological Measurement, 68*(4), 537–553.

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.

Schweizer, K. (2009). *Latent variable models representing the position effect and their application to reasoning items.* Paper presented at the Annual Meeting of the Psychometric Society, Cambridge, UK.

Shepard, L., Camilli, G., & Williams, D. (1984). Accounting for statistical artifacts in item bias research. *Journal of Educational Statistics, 9*(2), 93–128.

Wainer, H., & Skorupski, W. P. (2005). Was it ethnic and social-class bias or statistical artifact? Logical and empirical evidence against Freedle's method for reestimating SAT scores. *Chance, 18*(2), 17–24.