induced stereotype threat effects on test performance indeed resulted in measurement bias. They made use of a CFA model in which one latent factor (i.e., cognitive ability) was shown not to be able to explain group differences in test performance. In other words, they were able to detect measurement bias due to stereotype threat absent any additional information on, for instance, criterion performance. In fact, we happen to have data on the criterion performance (grade point average) of the subjects in Study 3 of Wicherts et al.; we tested for differential prediction but failed to find any (despite the presence of measurement bias).

The message that measurement bias does not necessarily result in underprediction is hardly new, but it has been largely ignored in the literature on selection fairness (Millsap, 2008). Given the potential social impact of measurement bias, this is an unfortunate state of affairs.

REFERENCES


The preparation of this comment was supported by VENI Grant 451-07-016 from the Netherlands Organization for Scientific Research (NWO) and Grant R01MH64707 from the National Institute of Mental Health. We thank Ryan P. Brown for kindly sharing his data.

Correspondence concerning this comment should be addressed to Jelte M. Wicherts, Department of Psychology, University of Amsterdam, Roetersestraat 15, 1018 WB, Amsterdam, The Netherlands. E-mail: j.m.wicherts@uva.nl

DOI: 10.1037/a0014924

Defense of Tests Prevents Objective Consideration of Validity and Fairness

Janet E. Helms
Boston College

In defending tests of cognitive abilities, knowledge, or skills (CAKS) from the skepticism of their “family members, friends, and neighbors” and aiding psychologists forced to defend tests from “myth and hearsay” in their own social networks (p. 215), Sackett, Borneman, and Connelly (May–June 2008) focused on evaluating validity coefficients, racial or gender group differences, and fairness assessment research. In doing so, they concluded that CAKS tests generally yield valid and fair test scores for their intended purposes, but because the authors did not adequately attend to (a) research design issues (e.g., inclusion of independent or predictor variables [IPV] and dependent variables or criteria), (b) statistical assumptions underlying interpretation of their analyses (e.g., bivariate normality of distributions of test scores and criteria), and (c) conceptual concerns (e.g., whether racial categories should be used as explanatory constructs), alternative conclusions about CAKS test score validity and fairness are plausible.

Although all of the foregoing areas of concern are germane to each of the assertions addressed by Sackett et al. (2008), the focus here is on Assertions 6 through 8 (p. 216; hereinafter called the fairness assertions [FA]) because making accurate inferences about fairness requires measurement experts to engage in a paradigmatic shift where sociodemographic groups (e.g., Blacks, Latinos/Latinas) are concerned, whereas, for the most part, addressing the other assertions merely requires a reminder of which standard psychometric principles have not been followed (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 1999).

The authors erroneously interpreted standardized mean differences (d) in CAKS test scores between Black and White test takers, for example, as if they were validity evidence, but they are not unless the intended purpose of tests is to disadvantage Blacks relative to Whites. Some of the ways in which d values differ from validity are as follows:

1. The d statistic merely describes the number of standard deviations (SDs) that separate two groups’ mean CAKS test scores rather than the actual test scores. Whereas correlations between test scores and criteria (i.e., validity coefficients) might be corrected for range restriction, measurement error, or whatever one can make a reasonable argument for correcting, Sackett et al. (2008) provided no rationale for correcting the numbers of SDs separating two groups’ CAKS test scores. In fact, Grissom and Kim (2001) argued that the correction for two groups’ SDs is to determine a priori whether the SDs are homogenous, whether they should be collapsed across groups to form the denominator of the d statistic and, therefore, make d values interpretable.

2. Racial/ethnic groups serve in the role of IPVs, a role that ordinarily is reserved for measures or manipulations of explanatory constructs, and CAKS test scores function in the role of dependent variables or criteria when d values are used to describe between-group racial/ethnic differences. In Sackett et al.’s (2008) analyses of validity coefficients, CAKS test scores served as measures of explanatory constructs (arguably, CAKS) for the selected CAKS criteria (e.g., grade point average). Yet the authors’ conclusion that racial groups’ d values reflected differences in “developed abilities” (p. 222) (a) treated the dependent variable (i.e., test scores) as explanatory and (b) therefore camouflaged the fact that the IPV in their FA research design was racial groups, a nominal variable with no conceptual meaning (Helms, Jernigan, & Mascher, 2005).

3. Sackett et al.’s (2008) interpretations of between-racial-group d values drew on “large-scale meta-analytic mean d values . . . [based on] . . . sample sizes of at least several hundred thousand” (p. 222). Large samples might be important in interpreting validity coefficients because they reduce error in estimating population values and perhaps minimize range restriction as an alternative interpretation of them. But d values are affected by disparities between the numbers of people in the two groups rather than total sample size per se. It is well known, for instance, that when the proportions of Blacks and Whites that are being compared on a continuous variable,
such as CAKS test scores, deviate from 50–50, the variability of the racial-group variable is restricted and the sizes of the resulting correlations (i.e., validity coefficients) and also $d$ values are adversely affected. Less obvious are the effects of disparate group sizes (i.e., skewed distributions of grouped test takers) on the between-racial-group mean differences and $SD$s from which $d$ values are calculated.

Racial-group skewness (RGS) levels are a critical concern in meta-analyses because samples are collapsed across studies to obtain $d$ values without considering group sizes, as was the case with respect to Sackett et al.’s (2008) arguments against the FA assertions. I calculated levels of skewness of White (coded 1) relative to Black (coded 0) test takers for the 40 samples obtained from 31 published studies reported in Helms (2008). In most of the samples (67.5%), racial groups, used as IPVs, were severely negatively skewed ($p < .001$), indicating that there were significantly more Whites than Blacks in the samples. The average CAKS test $d$ value was .99, which was consistent with Sackett et al.’s conclusions about the White–Black mean differences in CAKS test scores. The correlation between levels of RGS and $d$ values across studies was $r(40) = - .26$, $p < .11$ ($d = .43$); between mean differences and RGS, it was $r(40) = -.475$, $p < .002$ ($d = 1.08$); and between $SD$ and RGS, it was $r(40) = -.43$, $p < .004$ ($d = 0.99$).

These findings indicate that higher numbers of Whites relative to Blacks in samples were related to larger mean differences favoring Whites or disfavoring Blacks. They also indicate that the RGS correlations for both the numerator of the $d$ statistic (i.e., mean differences) and the denominator (i.e., aggregated $SD$s) were equivalent to about 1 $SD$ of test score variation; in fact, conversion of the latter RGS–$SD$ correlation to $d$ resulted in the same value as calculating $d$ by regular procedures (i.e., $d = 0.99$) and suggests that the average mean CAKS score differences and skewness levels across studies were interchangeable. In sum, when unequal or skewed racial groups are not corrected in the source studies used in meta-analyses, then the larger numbers of White relative to Black test takers provide an alternative explanation for what has been construed as a conceptually meaningful 1–$SD$ White–Black CAKS mean score difference. The atheoretical nature of racial groups makes it problematic that Sackett et al. (2008) interpreted the racial groups inherent in $d$ values as reflective of something more than mere relative frequency counts.

Explanation or removal of the variation in test scores ostensibly attributable to racial groups requires measurement of the types of constructs for which such groups serve as unacknowledged proxies (Helms, 2006; Helms et al., 2005). Most experts in race and cultural psychology, many of whom are cited in the *Guidelines on Multicultural Education, Training, Research, Practice, and Organizational Change* (APA, 2003), recognize that people (as opposed to tests) have measurable racial/cultural characteristics that interact with their responses to CAKS tests. Sackett et al. (2008) took exception to Steele and Aronson’s (1995) and Helms’s (2006) use of person-level conceptual racial variables to explain CAKS test scores.

Sackett et al.’s (2008) criticism of Steele and Aronson’s (1995) concept of stereotype threat (ST; i.e., fear of having negative societal stereotypes about one’s group confirmed with respect to oneself) was that although “motivational issues [i.e., ST] . . . may play a large role in the testing outcomes of individual test takers, [they cannot explain] group mean differences in high-stakes testing settings” (p. 225, italics added). Yet my calculation of the magnitude of the ST group effects using the $t$ values reported by Steele and Aronson (1995, p. 800) indicated that whether their analyses compared Blacks and Whites in the ST condition, $r(107) = 2.64, r = .25$, $d = 0.52$, or Blacks in ST and problem-solving conditions, $r(107) = 2.88, r = .27, d = 0.56$, or Blacks in ST and challenging-task conditions, $r(107) = 2.64, r = .25, d = 0.52$, the correlations were at least as large as the smallest of Sackett et al.’s uncorrected validity coefficients, which they interpreted as meaningful. Also, whereas ST did not account for 1 $SD$ of difference within conditions, the average effect (presuming equal groups) accounted for almost half of the averaged FA $d$ values for Blacks and Whites that Sackett et al. reported ($d_{FA} = .533/1.088 = .49$).

Helms (2006) showed that racial variables, such as racial identity schemas (i.e., self-perceptions derived from ascribed racial-group membership) can account for the equivalent of at least 1 $SD$ in individual Black test takers’ CAKS test scores, albeit by unknown mechanisms (pp. 851–852). Sackett et al. (2008) seemingly agreed that correlations between racial identity and test scores would indicate that tests assess construct-irrelevant variance but opined that “the dominant view in the testing field rejects the position that a finding of a relationship between race (or, in Helms’s model, a race-related variable) and test scores can be directly interpreted as signaling bias or unfairness” (p. 223) because racial identity might also be associated with “investing less time in one’s studies” (p. 223), which the authors interpreted as construct-relevant variance (i.e., CAKS). Yet if study time is also correlated with CAKS test scores, then it is another source of construct-irrelevant variance that contributes to unfairness (AERA et al., 1999). Also, when researchers directly interpret racial groups, which are not measurable, as if they are IPVs, they are engaging in unfair assessment, a problem that can be overcome by replacing racial-group frequency counts with “race-related” IPVs, such as ST or racial identity, which can be measured. This type of paradigm shift in FA research is long past due.

REFERENCES


Correspondence concerning this comment should be addressed to Janet E. Helms, Institute for the Study and Promotion of Race and Culture, Boston College, 318 Campion Hall, Chestnut Hill, MA 02467. E-mail: helmsja@bc.edu