focused on the technical arena, and they acknowledged that other questions (i.e., values-based) should be considered. What was not mentioned, however, is that these questions are not mutually exclusive; how one defines the criteria will ultimately impact the technical qualities of a test.

In thinking about expanding the criteria, consider creativity, one of many noncognitive factors proposed to supplement high-stakes testing. Enright and Gitomer’s (1989) study identified creativity as one of the desired admissions competencies. Creativity is also considered crucial to both local and global economic success (Florida, 2002). People who are creative, in addition to being happier and in better physical health, are also more likely to be a successful entrepreneur, to rise in a company, to persevere, and to produce better dissertations (for an overview, see Kaufman & Beghetto, in press). Yet measures of individual differences in creativity do not predict many traditional school or work indicators because such criteria often fail to incorporate creative or innovative performance (Sterngart & Williams, 1997).

Why would creative achievements be a good choice as an additional criterion? In addition to their value as performance outcomes, measures of creativity show few differences across gender or ethnicity. In addition, self-perceptions of creativity are often the mirror image of self-perceptions of cognitive ability. One study found that African Americans and Native Americans had markedly higher self-assessments of their creative abilities than their peers (Kaufman, 2006). Therefore, testing contexts that incorporate less stereotypically threatening tasks may reduce test anxiety in underperforming groups.

Despite arguments for the inclusion of creative performance in school- and work-based criteria, creativity assessment is inconsistent, often time-consuming, and expensive. There are valid reasons why organizations and schools may not choose to tackle such an undefined topic. Such efforts are merit worthy, however, and one example has proved quite fruitful. Sterngart (2008) changed the admission procedures at Tufts University so that all components of his successful intelligence model (including creativity) are optional predictors. He found that this new system predicts college success more accurately than standard admissions tests alone; in addition, ethnic differences are significantly reduced. The quality of applicants rose (indeed, despite deemphasizing SAT scores, the average SATs of applicants increased), and minority admissions went up (Sterngart, 2008).

Popular opinions are often misinformed. Sometimes, however, these opinions shed light on the limitations of an accepted approach. We do not disagree with the technical points that Sackett et al. (2008) argued. We believe, however, that the popular opinion that high-stakes tests do not always predict real-life success may reflect that our efforts to measure performance are lacking. We offer creativity as an example of one of many noncognitive characteristics that could be used both as a way of identifying new criteria and/or as a component of the selection process. We are not arguing the importance of high-stakes tests; we just think the entire process could benefit from an expanded consideration of creativity.

REFERENCES


Correspondence concerning this comment should be addressed to James C. Kaufman, Department of Psychology, California State University, San Bernardino, 5500 University Parkway, San Bernardino, CA 92407-2318. E-mail: jkaufman@csusb.edu

DOI: 10.1037/a0014992

The Absence of Underprediction Does Not Imply the Absence of Measurement Bias

Jelte M. Wicherts
University of Amsterdam

Roger E. Millsap
Arizona State University

Sackett, Borneman, and Connelly (May–June 2008) recently discussed several criticisms that are often raised against the use of cognitive tests in selection. One criticism concerns the issue of measurement bias in cognitive ability tests with respect to specific groups in society. Sackett et al. (2008) stated that “absent additional information, one cannot determine whether mean differences [in test scores] reflect true differences in the developed ability being measured or bias in the measurement of that ability” (p. 222). Their discussion of measurement bias appears to suggest that measurement bias in tests can be accurately detected through the study of differential prediction of criteria across groups. In this comment, we argue that this assertion is incorrect. In fact, it has been known for more than a decade that tests of differential regression are not generally diagnostic of measurement bias (Millsap, 1997, 1998, 2008).

Differential prediction implies differences across groups in the prediction of criterion scores (e.g., grade point average, ratings of job performance) from ability test scores. Differential prediction can be revealed in the regression context by group differences in the regression lines relating the criterion scores to the ability test scores. Measurement bias exists when two individuals who are identical on the construct(s) measured by a test but who are from different groups have different probabilities of attaining the same score on the test (i.e., they have different expected test scores). A test is considered free of measurement bias, or measurement invariant, if the two persons described above have the same probability of attaining any score on the test (Mellenbergh, 1989). Sackett et al. (2008) subscribe to this definition of measurement invariance, as do we. Measurement invariance in the test can be studied directly at the item, parcel, or subtest level by adopting measurement models such as those from item response theory or confirmatory factor analysis (CFA). Within these measurement models, the equality over groups...
of parameters that relate latent variables to test scores can be tested statistically (Meredith, 1993; Millsap & Everson, 1993). On the other hand, the demonstration of identical test-criterion regressions across groups is not sufficient to establish measurement invariance. For example, it is easily shown that under a common factor model for the test and criterion, measurement bias can be manifested in group differences in measurement intercepts even if the regression of the criterion on the test is identical across groups (Millsap, 2008).

Sackett et al. (2008) discussed the potentially biasing effects of stereotype threat on performance on cognitive ability tests. Stereotype threat refers to the pressure that minority test takers may experience when confronted with a negative stereotype related to their group’s ability. In several experimental studies, stereotype threat has been shown to negatively affect the test performance of minorities. Sackett et al. asserted that the biasing effects of stereotype threat on minority test scores must result in differential prediction (underprediction) of criterion performance of minorities. In light of the points we have raised here, Sackett et al.’s assertion is not true.

To illustrate this, let us consider data from Brown and Day (2006), who experimentally tested the effects of stereotype threat on performance on Raven’s Advanced Progressive Matrices (APM) test among Black students. These Black students also provided their ACT scores, as did a group of White students who acted as a control group. These data enable a study of differences between the White and Black groups in the prediction of ACT scores by the APM scores. The results are depicted in Figure 1. The top scatterplot is based on the APM scores from the condition in which the effects of stereotype threat were absent, whereas the bottom scatterplot is based on the scores from the (two) conditions in which stereotype threat effects were found to depress scores of Blacks on the APM. The top figure shows the so-called overprediction effect for the Black group; if we were to predict the ACT scores in the Black group by means of either the common regression line (not depicted) or the (dashed) regression line based on the White group, the predicted ACT scores of the Black students would be higher than their actual ACT scores. Such overprediction effects are found empirically in other contexts as well. Note that the upper scatterplot has no necessary implication for measurement bias with respect to Whites (Millsap, 2008). The bottom scatterplot shows a less severe case of overprediction. Results like those in the bottom scatterplot are often proffered in support of the argument that tests are not biased in a measurement sense with respect to the Black group. In truth, however, Brown and Day experimentally lowered the scores of Blacks in the conditions that gave rise to the bottom scatterplot. These scores are strongly biased with respect to the Blacks, but the differential prediction analysis appears to suggest that the test is nearly unbiased in the measurement sense, following Sackett et al.’s (2008) reasoning.

Sackett et al.’s (2008) assertion that measurement bias necessarily results in underprediction of criterion performance is false. Measurement invariance should be tested directly by using psychometric models. Direct tests of this sort can support conclusions from attempts to experimentally induce measurement bias. For example, Wicherts, Dolan, and Hessen (2005) found in three studies that experimentally...
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, Roetersstraat 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 skeptical social networks (p. 215), Sackett, Borneman, and Connelly (May–June 2008) focused on evaluating validity coefficients, racial or gender group differences, and fair 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 [IPVs] 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 homogeneous, 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,