Running head: MEASUREMENT BIAS AND DIFFERENTIAL PREDICTION

The absence of underprediction does not imply the absence of measurement bias

Jelte M. Wicherts

University of Amsterdam

Roger E. Millsap

Arizona State University

This article has been accepted for publication in American Psychologist

Jelte M. Wicherts, Department of Psychology, University of Amsterdam, Roetersstraat 15, 1018 WB, Amsterdam, The Netherlands, j.m.wicherts@uva.nl. The preparation of this article was supported by VENI grant no. 451-07-016 from the Netherlands Organization for Scientific Research (NWO). The authors would like to thank Ryan P. Brown for kindly sharing his data.

Sackett, Borneman, & Connely (2008) recently discussed several criticisms pertaining to 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. state 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., GPA, 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 above have the same probability of attaining any score on the test (Mellenbergh, 1989). Sackett et al. 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 (IRT) 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. discuss the potentially biasing effects of stereotype threat on cognitive ability tests performance. 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 test performance of minorities. Sackett et al. assert 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 above points, Sackett et al.'s assertion is simply 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 will be higher than their actual ACT scores. Such overprediction effects are found

empirically in other contexts as well. 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 the test is 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 reasoning.

Sackett et al.'s 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 criterion performance (GPA) of the subjects in Study 3 of Wicherts et al., 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

Brown, R. P., & Day, E. A. (2006). The difference isn't Black and White: Stereotype threat and the race gap on Raven's Advanced Progressive Matrices. *Journal of Applied Psychology, 91,* 979-985.

Mellenbergh, G. J. (1989). Item bias and item response theory. *International Journal of Educational Research, 13,* 127-143.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58,* 525-543.

Millsap, R. E. (1997). Invariance in measurement and prediction: Their relationship in the single-factor case. *Psychological Methods, 2,* 248-260.

Millsap, R.E. (1998). Group differences in regression intercepts: Implications for factorial invariance. *Multivariate Behavioral Research, 33,* 403-424.

Millsap, R. E. (2008). Invariance in measurement and prediction revisited. *Psychometrika, 72,* 461-473.

Millsap, R. E., & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement, 17,* 297-334.

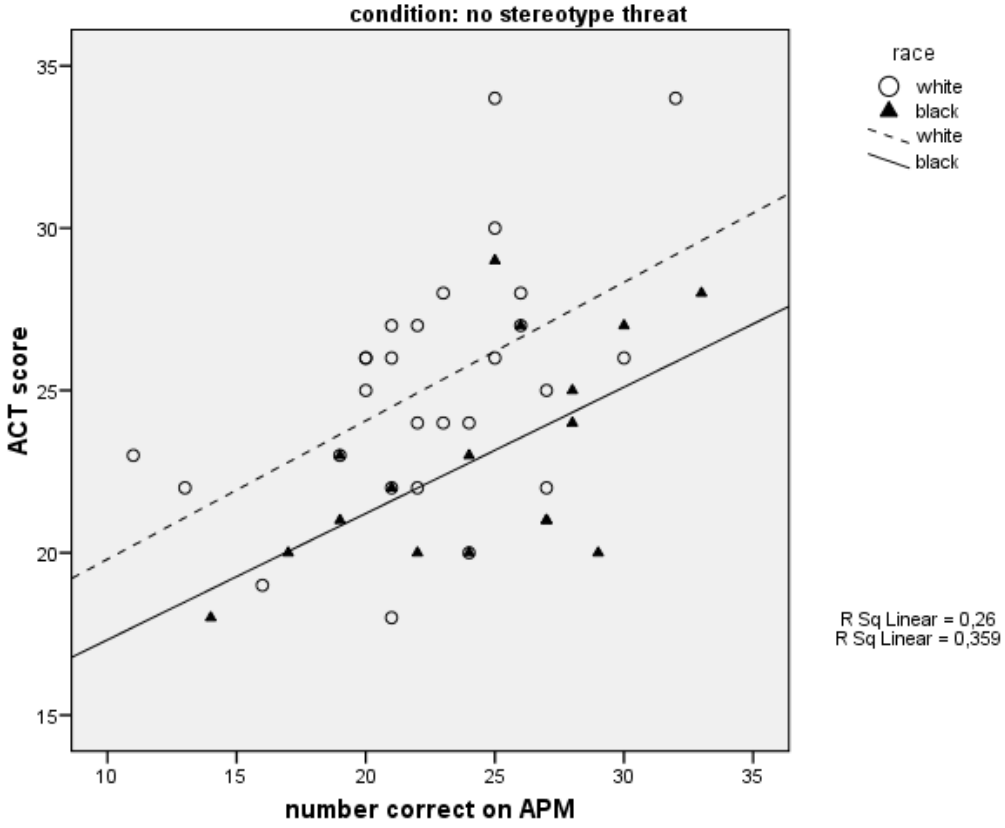
Sackett, P. R., Borneman, M. J., & Connelly, B. S. (2008). High-stakes testing in higher education and employment: Appraising the evidence for validity and fairness. *American Psychologist, 63,* 215-227.

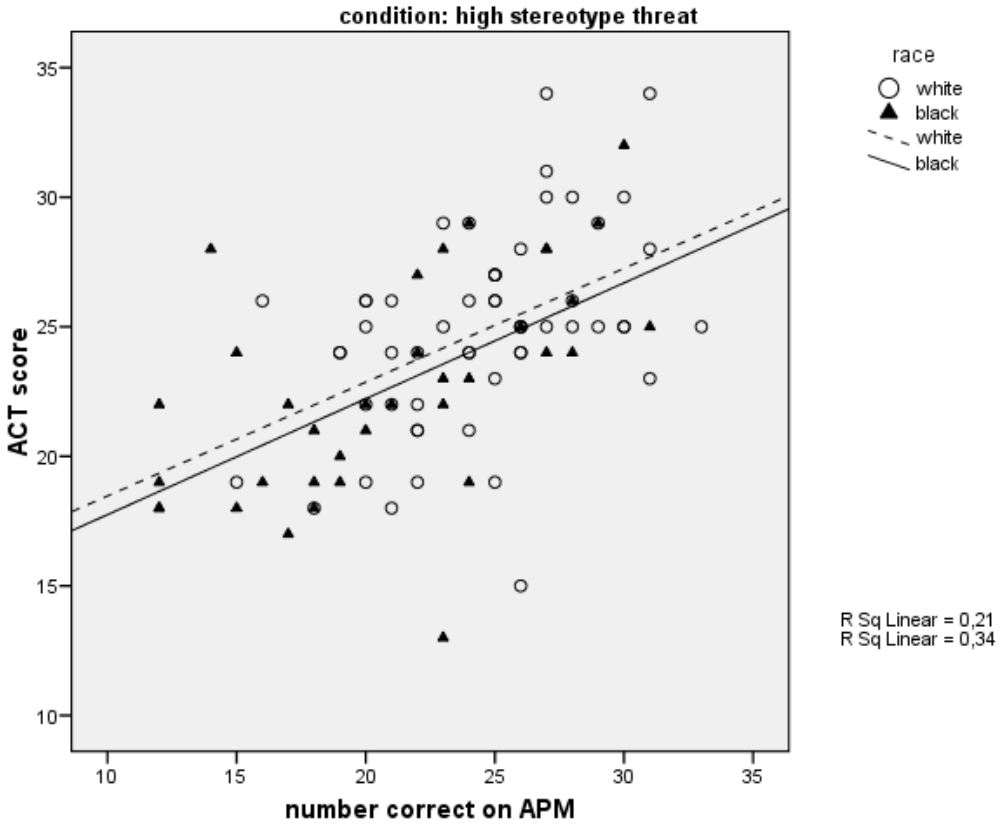
Wicherts, J. M., Dolan, C. V., & Hessen, D. J.(2005). Stereotype threat and group differences in test performance: A question of measurement invariance. *Journal of Personality and Social Psychology, 89,* 696-716.

Figure Caption

Figure 1.

Regression lines for the prediction of ACT scores with Advanced Progressive Matrices for Blacks and Whites in conditions without stereotype threat (top) and conditions with stereotype threat (bottom) from Brown & Day (2006).


condition: no stereotype threat
race
white
black
white
black
ACT score
number correct on APM
R Sq Linear = 0,26
R Sq Linear = 0,359


condition: high stereotype threat
race
white
black
white
black
ACT score
number correct on APM
R Sq Linear = 0,21
R Sq Linear = 0,34