role in the race gap are understood in this literature and related material. A broader review would give them much less reason for concern.

Second, Sackett et al.'s (2004) narrow focus may have also led them to worry too much about the use of covariance analysis in Steele and Aronson's (1995) study. They worried that this analysis led readers to believe that African Americans performed as well as Whites in the nondiagnostic (no stereotype threat) condition of that experiment, when, in fact, without this adjustment, they would be shown to perform still worse than Whites, as predicted by the group difference in their SATs. We, as much as Sackett et al., regret any confusion that this common analysis may have caused. We used it to reduce error variance and thus make the experiment more sensitive to the effect of conditions, especially in light of our small number of participants.

But again, the larger stereotype threat literature is critical. It shows the effect of stereotype threat on an array of tests—SATs, IQ tests, and French language tests to list only a few—sometimes with a covariance adjustment, but many times without. Whatever impression readers got from the use of covariance in Steele and Aronson (1995) would certainly have been corrected by this larger literature. They would know (a) that the skills measured by the SAT can indeed affect subsequent test performance, (b) that under common and important conditions, stereotype threat has powerful effects of its own on test performance, and (c) that detecting an effect of stereotype threat on test performance does not depend on the use of covariance analysis.

We note here that even in Study 2 of Steele and Aronson (1995), the effect of stereotype threat does not depend on the use of the SAT covariate. African Americans in the diagnostic (stereotype threat) condition performed a full standard deviation lower than African Americans in the nondiagnostic (no threat) condition—a 3-item effect on a 26-item test that was significant without the use of a covariate. Also, the interaction that tested whether the effect of stereotype threat was greater for African Americans than for Whites reached a one-way level of significance, $F = 3.75$, $p < .06$, with no covariate and only 10 participants per cell.

Third, Sackett et al. (2004) stated that

> absent stereotype threat, the African American–White difference is just what one would expect based on the African American–White difference in SAT scores, whereas in the presence of stereotype threat, the difference is larger than would be expected based on the difference in SAT scores. (p. 9)

They seem to be saying that the nondiagnostic (no stereotype threat) condition embodied the conditions of regular testing because it reproduced the African American–White difference observed on the regular SAT (i.e., no mean difference once adjusted for SATs) and that the diagnostic condition imposed an extra threat not typical of regular testing because it caused African Americans to perform worse than their SATs would have predicted.

However, seeing the pattern of African American–White differences in the nondiagnostic condition as more "expected" from SATs is, we believe, over-reading the data. The Graduate Record Examination (GRE) is correlated with the SAT, but not perfectly. And recall our small number of participants. Under these conditions—even under better conditions—SATs could not predict GREs so precisely. Thus, one cannot say which of the two African American–White differences—the threat difference or the no-threat difference—is best expected from the group difference in SATs, let alone which of the two conditions is most like regular testing.

Again, the larger literature is relevant. There (as in Steele & Aronson, 1995) it is the stereotype threat conditions, and not the no-threat conditions, that produce group differences most like those of real-life testing. Stereotype threat conditions represent the test as ability diagnostic, either en passant or by saying nothing at all and relying on participants to know a test when they see one. It is the no-threat conditions that are unlike real-life testing. They present the test as nondiagnostic of the participants' ability or of their group's ability—in stark contrast to real-life testing situations. Yet it is the stereotype threat conditions that impair performance among the people who are subject to being negatively stereotyped (African Americans in the case of the Steele and Aronson experiments). The big picture, then, rather than guesses based on the pattern of results in a single experiment, should be used to judge which of these conditions—stereotype threat or no stereotype threat—is most like real-life testing.

Twenty-nine mischaracterizations of any research finding are 29 too many. However, using the frequency of these mischaracterizations to signal concern, while ignoring the large amount of information that would allay that concern, only furthers misunderstanding. Sackett et al. (2004) ignored the large number of discussions in the relevant literatures and media reports that do not overattribute the race gap to stereotype threat—discussions that vastly outnumber 29. Thus, rather than these mischaracterizations constituting a gathering danger, they are just mischaracterizations, almost completely ignored and having whatever misunderstanding they do cause constantly corrected by the natural progress of research.

REFERENCES

Aronson, J., Lustina, M., Good, C., Keough, K., Steele, C. M., & Brown, J. (1999). When White men can't do math: Necessary and sufficient factors in stereotype threat. *Journal of Experimental Social Psychology, 35,* 29–46.

Bowen, W. G., & Bok, D. (1998). *The shape of the river: Long-term consequences of considering race in college and university admissions.* Princeton, NJ: Princeton University Press.

Jencks, C., & Phillips, M. (Eds.). (1998). *The Black–White test score gap.* Washington, DC: Brookings Institution Press.

Massey, D. S., Charles, C. Z., Lundy, G. F., & Fischer, M. J. (2003) *The source of the river: The social origins of freshmen at America's selective colleges and universities.* Princeton, NJ: Princeton University Press.

Sackett, P. R., Hardison, C. M., & Cullen, M. J. (2004). On interpreting stereotype threat as accounting for African American–White differences on cognitive tests. *American Psychologist, 59,* 7–13.

Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology, 69,* 797–811.

Correspondence concerning this comment should be addressed to Claude M. Steele, Department of Psychology, Stanford University, Jordan Hall, Building 420, Stanford, CA 94305-2130. E-mail: steele@psych.stanford.edu

DOI: 10.1037/0003.066X.59.1.48

On the Value of Correcting Mischaracterizations of Stereotype Threat Research

Paul R. Sackett, Chaitra M. Hardison, and Michael J. Cullen
University of Minnesota, Twin Cities Campus

We see no disagreement by Steele and Aronson (2004, this issue) with the key issues that prompted our article (Sackett, Hardison, & Cullen, 2004, this issue). They

agree that it is a misinterpretation of the Steele and Aronson (1995) results to conclude that eliminating stereotype threat eliminates the African American–White test-score gap. They agree that we have identified multiple mischaracterizations of their work in media reports, journal articles, and textbooks, which wrongly interpret their work as finding that eliminating stereotype threat did indeed eliminate the score gap. They agree that these mischaracterizations are regrettable.

However, Steele and Aronson (2004) assert that there is no need to worry about mischaracterizations of their findings in the absence of evidence that these mischaracterizations have led to widespread misunderstanding of the role stereotype threat plays in explaining the African American–White test-score gap. We disagree. Although evidence of such misunderstanding would certainly be grave cause for concern, we believe it is sufficiently worrisome when one of the seminal studies on stereotype threat is commonly wrongly interpreted—by the popular media, textbook publishers, and academics alike—to mean that the African American–White test-score gap disappears when stereotype threat is eliminated. Steele and Aronson assert that their 1995 study is "a drop in an ocean of information about the race gap" (Steele & Aronson, 2004, p. 47). We believe they are unduly modest about the impact of their paper; that the Social Sciences Citation Index reports that it has been cited more than 300 times is one indicator of its prominence.

Steele and Aronson (2004) assert that because there are now over 100 research studies on stereotype threat, our focus on the first article on the topic results in a serious bias. However, they later acknowledge that their article is one of few stereotype threat studies focusing on African Americans. As the African American–White score gap was the topic of our article, we see our focus on this pivotal and highly cited article as entirely appropriate.

Steele and Aronson (2004) also assert that no attentive reader of the literature on the race gap would conclude that stereotype threat is its sole cause. However, our concern is with broader audiences than the serious scholar working on issues of race. We are concerned about students who are being initially exposed to issues of psychological testing and the race gap in their introductory psychology courses. We are concerned about managers responsible for personnel selection systems in their organizations. We are concerned about psychologists who do not follow testing issues closely and whose only exposure to stereotype threat may be through an American Psychological Association *Monitor on Psychology* column making the interpretive error that is the focus of our article. We are concerned about the large audience watching *Frontline* and hearing that the score gap is eliminated in the no-threat condition.

Steele and Aronson (2004) address the use of a prior SAT score as a covariate, claiming that we overworried about readers being misled by this analysis. They argue that a larger literature shows the stereotype threat effect, sometimes with the use of a prior test as a covariate and sometimes without. However, in our article, we noted clearly that we are not questioning the finding of a stereotype threat effect (i.e., the finding of a Race × Diagnostic Condition interaction) in Steele and Aronson (1995). Our concern is with misinterpreting the graphical presentation of findings as suggesting that group differences can be eliminated.

Steele and Aronson (2004) take issue with our comparison of African American–White differences on the prior SAT and on GRE-based scores in the two experimental conditions. Steele and Aronson assert that these are not comparable because the pretest SAT and the experimental GRE-based test are not perfectly correlated and because N is small. Given the extensive data on the similarity of the score gaps between the two tests and the correlation between the two, we see it as reasonable to posit that two groups that do not differ on the SAT would also be expected not to differ on the GRE.

We share with Steele and Aronson the beliefs that single experiments do not answer all questions and that it is important to examine the role of stereotype threat in real-life testing settings. We certainly agree with their position that evolving literatures have self-correcting capacities, and we view our article as fulfilling exactly such a role. Most crucially, we note that the disagreement between us is about the consequences of the mischaracterization we documented, not about whether the work has been mischaracterized.

REFERENCES

Sackett, P. R., Hardison, C. M., & Cullen, M. J. (2004). On interpreting stereotype threat as accounting for African American–White differences on cognitive tests. *American Psychologist, 59,* 7–13.

Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology, 69,* 797–811.

Steele, C. M., & Aronson, J. (2004). Stereotype threat does not live by Steele and Aronson (1995) alone. *American Psychologist, 59,* 47–48.

Correspondence concerning this comment should be addressed to Paul R. Sackett, Department of Psychology, University of Minnesota, Elliott Hall, 75 E. River Road, Minneapolis, MN 55455. E-mail: psackett@tc.umn.edu

DOI: 10.1037/0003.066X.59.1.49

Journal Impact Factors and Self-Citations: Implications for Psychology Journals

Frederik Anseel, Wouter Duyck, and Wouter De Baene
Ghent University

Marc Brysbaert
Royal Holloway University of London

Recently, Adair and Vohra (January 2003) analyzed changes in the number of references and citations in psychology journals as a consequence of the current knowledge explosion. In their study, the authors made a striking observation of the sometimes excessive number of self-citations in psychology journals. However, after this illustration, no further attention was paid to the issue of self-citation. This is unfortunate because little is known about self-citing practices in psychology. Early research on self-citations in psychology journals indicated that about 10% of citations were self-citations, and one author concluded that "it is apparent that controlling for self-citation is not necessary" (Gottfredson, 1978, p. 932). Similarly, although the *Publication Manual of the American Psychological Association* (American Psychological Association, 2001) provides clear guidelines on the form citations should take, it does not indicate when it is appropriate to cite one's own work.

Recent figures urge more caution when dealing with self-citations. A multidisciplinary study found that 36% of all citations represent author self-citations (Aksnes, 2003; see also McGarty, 2000, for a similar finding in social psychology). Especially troublesome is the finding that self-citations peak during the first three years after publication, thereby strongly influencing impact factors of journals that are based on two-year periods.

Although the use of citation counts (and impact factors) has been criticized in all disciplines (see, e.g., Boor, 1982), it has become the main quantitative measure of the quality of a journal. Accordingly, these figures are used to make decisions about