On Interpreting Stereotype Threat as Accounting for African American–White Differences on Cognitive Tests

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

C. M. Steele and J. Aronson (1995) showed that making race salient when taking a difficult test affected the performance of high-ability African American students, a phenomenon they termed stereotype threat. *The authors document that this research is widely misinterpreted in both popular and scholarly publications as showing that eliminating stereotype threat eliminates the African American–White difference in test performance. In fact, scores were statistically adjusted for differences in students' prior SAT performance, and thus, Steele and Aronson's findings actually showed that absent stereotype threat, the two groups differ to the degree that would be expected based on differences in prior SAT scores. The authors caution against interpreting the Steele and Aronson experiment as evidence that stereotype threat is the primary cause of African American–White differences in test performance.*

Mean differences in test scores between various racial/ethnic groups are commonly observed when tests of knowledge, skill, ability, or achievement are used in education and employment contexts. A large amount of research has been devoted to attempting to understand the causes of these mean differences and to ameliorating them (see, e.g., Bobko, Roth, & Potosky, 1999; Hartigan & Wigdor, 1989; Pulakos & Schmitt, 1996; Sackett & Ellingson, 1997; see Sackett, Schmitt, Ellingson, & Kabin, 2001, for a review). The test-score gap remains one of the most pressing societal issues of the day. It is an issue that is not confined to discussion among psychologists and psychometricians; few issues in psychology attract as much attention from the general public. Consider, for example, the amount of public attention received by *The Bell Curve* (Herrnstein & Murray, 1994) upon its publication in 1994. In recent years, the theory of stereotype threat (Steele & Aronson, 1995) has received a great amount of scientific and popular attention as a potential contributor to mean differences in test scores. Although the term was first introduced into the literature only in 1995, stereotype threat is covered in two thirds of a sample of current introductory psychology textbooks that we describe later in this article, indicating extraordinarily rapid incorporation of the concept into the psychological mainstream.

Steele and colleagues hypothesized that when a person enters a situation in which a stereotype of a group to which the person belongs becomes salient, concerns about being judged according to that stereotype arise and inhibit performance. Although this phenomenon can affect performance in many domains, one area that has been the focus of much research is the applicability of stereotype threat to the context of cognitive ability testing. According to the theory, when members of racial minority groups encounter tests, their awareness of the common finding that members of some minority groups tend to score lower on average on tests leads to concern that they may do poorly on the test and thus confirm the stereotype. This concern detracts from their ability to focus all of their attention on the test and results in poorer test performance. Similar effects have been hypothesized for gender in the domain of mathematics, where stereotypes that women do not perform as well as men are common. A boundary condition for this is proposed, namely, that individuals identify with the domain in question. If competence in a domain (e.g., mathematics) is something with which the individual identifies, stereotype threat will be experienced. If the domain is not relevant to the individual's self-image, the testing situation will not elicit stereotype threat.

Steele and Aronson (1995) initially obtained support for this theory through a series of laboratory experiments. The basic paradigm is to use high-achieving majority and minority students as research participants and compare test performance when stereotype threat is induced and when it is not. One mechanism for inducing threat is via instructional set. In the stereotype threat condition, participants are told that they will be given a test of intelligence; in the nonthreat condition, they are told they will be given a problem-solving task that the researchers have developed. In fact, all participants receive the same test. Steele and Aronson reported a larger majority–minority difference in the threat condition than in the nonthreat condition, a finding supportive of the idea that the presence of stereotype threat inhibits minority group performance. This find-

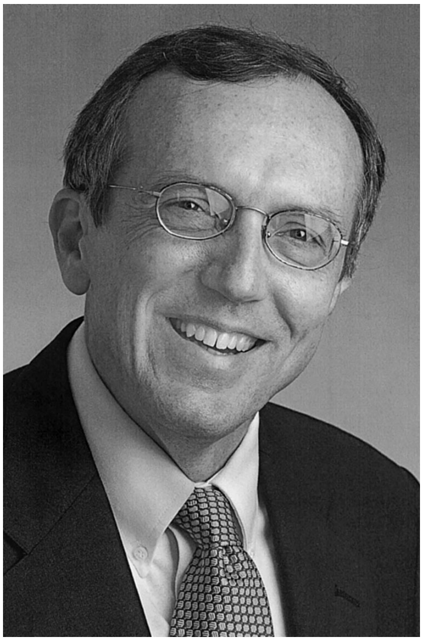
Editor's note. Lloyd Bond served as action editor for this article.

Author's note. Paul R. Sackett, Chaitra M. Hardison, and Michael J. Cullen, Department of Psychology, University of Minnesota, Twin Cities Campus.

We thank the College Board for supporting this research.

Correspondence concerning this article 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

Copyright 2004 by the American Psychological Association, Inc. 0003-066X/04/$12.00
Vol. 59, No. 1, 7–13 DOI: 10.1037/0003-066X.59.1.7



Paul R. Sackett

ing is well replicated (Aronson et al., 1999; Quinn & Spencer, 1996, 2001; see Steele, Spencer, & Aronson, 2002, for a review). In some settings, the threat-inducement mechanism is simply asking participants to indicate their race prior to taking the test; this alone is enough to induce stereotype threat in these lab settings (Croizet & Claire, 1998; Shih, Pittinsky, & Ambady, 1999; Steele & Aronson, 1995).

One question that quickly arises from this research is the degree to which this phenomenon generalizes from the laboratory to applied settings, such as admissions testing for higher education and employment testing, though only a few studies to date have examined threat in applied testing settings (Cullen, Hardison, & Sackett, in press; Stricker & Ward, in press). Some have interpreted the Steele and Aronson (1995) findings as indicating that majority–minority test-score differences are due solely to stereotype threat: If not for the presence of stereotype threat, scores for majority and minority groups would be comparable. Here are two examples. First, in the fall of 1999, the PBS show *Frontline* broadcast a one-hour special entitled "Secrets of the SAT" (Chandler, 1999), in which stereotype-threat research was featured. The research was described by the program's narrator as follows:

> At Stanford University, psychology professor Claude Steele has spent several years investigating the 150-point score gap[1] between Whites and Blacks on standardized tests. Was the cause class difference, lower incomes, poorer schools, or something else? . . . In research conducted at Stanford, Steele administered a difficult version of the Graduate Record Exam, a standardized test like the SAT. To one set of Black and White sophomores, he indicated that the test was an unimportant research tool, to other groups that the test was an accurate measure of their verbal and reasoning ability. Blacks who believed the test was merely a research tool did the same as Whites. But Blacks who believed the test measured their abilities did half as well. Steele calls the effect "stereotype threat." (Chandler, 1999)

Note that this description suggests that the "150-point score gap" was eliminated when stereotype threat was eliminated ("Blacks who believed the test was merely a research tool did the same as Whites," Chandler, 1999).

Second, the American Psychological Association's then-Executive Director for Science, Richard McCarty, devoted his April 2001 *Monitor on Psychology* column to Steele's work. McCarty (2001) correctly characterized Steele's work as showing that African American students scored lower on a test when it was labeled a measure of intelligence than when it was not given that label. More importantly, he asserted that when the test was not labeled as a measure of intelligence, African American students performed just as well as White students.

However, McCarty (2001) and *Frontline* (Chandler, 1999) failed to note that Steele's work examined African American and White students statistically equated on the basis of prior SAT scores. What Steele and Aronson (1995) reported was not that actual test scores were the same for African American and White students when threat was removed but rather that after scores were statistically adjusted for differences in students' prior SAT performance, scores of both groups were the same. Thus, the findings actually show that absent stereotype threat due to labeling the test as a measure of intelligence, the African American and White students differed to about the degree that would be expected on the basis of differences in prior SAT scores.[2]

To understand why this is critical, consider Figure 1. Figure 1A is a reproduction of the key findings from Steele and Aronson's (1995) original study; this graph is frequently reproduced in presentations for broader audiences, such as Steele's (1997) *American Psychologist* article and Steele and Aronson's (1998) contribution to Jencks and Phillips's (1998) book on the African American–White score gap. Visually, one sees an African American–White gap in the threat condition and no gap in the no-threat condition. The dependent variable is labeled "Mean items solved, adjusted by SAT." Thus, although Steele and Aron-

[1] We are unable to explain the *Frontline* program's reference to a "150-point score gap between Whites and Blacks on standardized tests" (Chandler, 1999). The measuring scales for standardized tests vary widely, and thus, the magnitude of the African American–White gap will be different for each test. To deal with the use of different scales for different tests, it is common to express the score gap in terms of the standardized mean difference (*d*) between the groups: the difference between the means divided by the standard deviation. The prototypic finding is a 1.0 standard deviation difference between the two groups.

[2] African American and White students in this sample equated on prior SAT scores would not be expected to score exactly the same on a subsequent test unless both groups were drawn from populations with the same mean. It is likely that the two groups were drawn from populations with different SAT means, and thus, scores would be expected to regress toward the mean for the respective groups. Regression effects should be very small, though, as there is no reason to expect the students participating in the study to differ substantially from the African American and White means for the population of Stanford students, whereas regression effects are meaningful at the extremes of distributions.

Chaitra M. Hardison

son have been clear about the fact that participants are equated on the basis of initial SAT scores, it is not clear that the implications of this will be grasped by the reader.

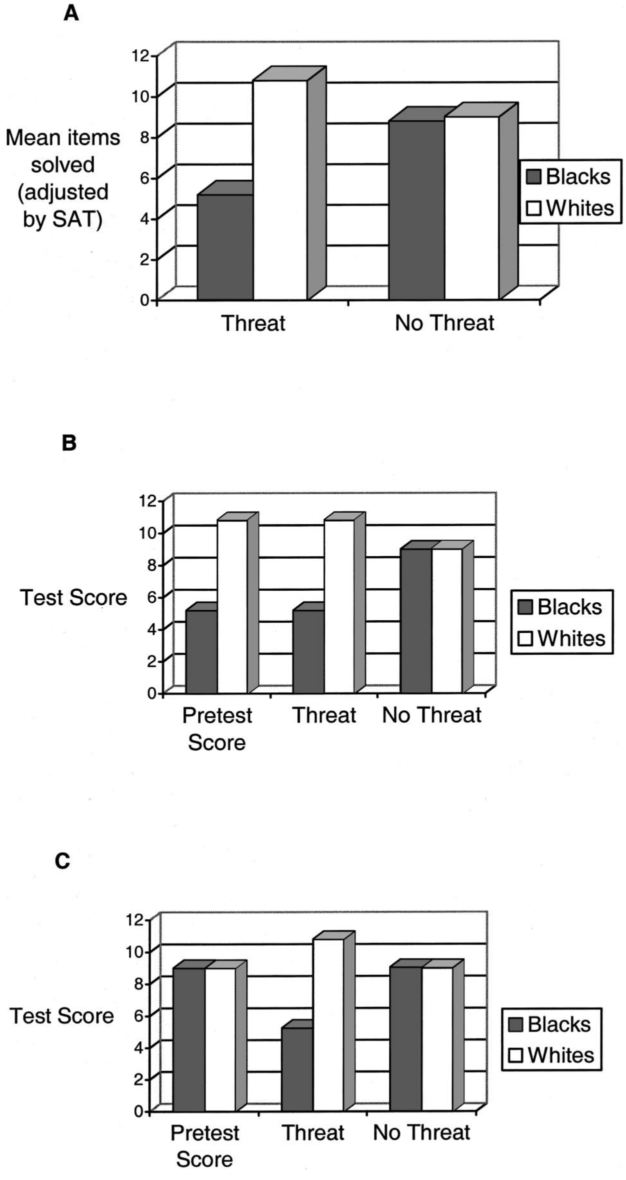
Figure 1B is our characterization of what we believe is implicitly assumed by many readers when they confront Figure 1A in reading Steele and Aronson's work. We have added a condition to the graph, namely, the commonly observed African American–White difference on tests like the GRE and the SAT. Readers may implicitly add to Figure 1A their knowledge about this commonly observed gap and interpret the research as follows: "There is a large score gap on commonly used tests; this mirrors the gap found in the threat condition in Steele and Aronson's work. But when threat is eliminated, the gap disappears." In other words, eliminating threat eliminates preexisting differences.

This interpretation is incorrect. Figure 1C is our characterization of the appropriate way to interpret Steele and Aronson's (1995) work. Here, we have also added a condition to the graph, reflecting the equating of the two groups in terms of their performance on the SAT. Figure 1C can be interpreted as follows: "In the sample studied, there are no differences between groups in prior SAT scores, as a result of the statistical adjustment. Creating stereotype threat produces a difference in scores; eliminating threat returns to the baseline condition of no difference." This casts the work in a very different light: Rather than suggesting stereotype threat as the explanation for SAT differences, it suggests that the threat manipulation creates an effect independent of SAT differences.

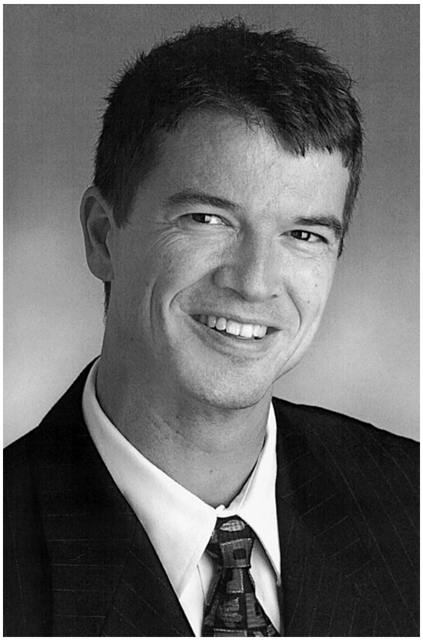
Thus, rather than showing that eliminating threat eliminates the large score gap on standardized tests, the research actually shows something very different. Specifically, absent stereotype threat, the African American–White difference is just what one would expect based on the African American–White 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.

It is important to note that this is a misinterpretation made by McCarty (2001) and by *Frontline* (Chandler, 1999), not by Steele and Aronson (1995) in their original

Figure 1
Interpretations of Steele and Aronson's Findings



Note. Figure 1A is an adaptation of Figure 2 from "Stereotype Threat and the Intellectual Test Performance of African Americans," by C. M. Steele and J. Aronson, 1995, *Journal of Personality and Social Psychology, 69,* p. 802. Copyright 1995 by the American Psychological Association. Adapted with permission of the authors.



Michael J. Cullen

research: The graphs in Steele and Aronson's research that document score differences consistently label them "adjusted by SAT." It is also important to note that the above observations are not meant as criticisms of Steele and Aronson's research. Steele and Aronson clearly demonstrated a very interesting phenomenon in a series of persuasive and carefully conducted experiments. They have shown that stereotype threat can affect the performance of some students on some tests, an important finding worthy of careful exploration. What they have not done, and do not purport to do, is to offer stereotype threat as the general explanation for the long-observed pattern of subgroup differences on a broad range of cognitive tests. Our concern, though, is that others (e.g., *Frontline*) do, in fact, interpret the research as implying that stereotype threat plays a broader explanatory role for subgroup differences.

Extent of Misinterpretation of Steele and Aronson

In the presentation above, we have focused on two specific cases in which the failure to recognize the implications of the statistical adjustment for existing SAT differences led to the incorrect conclusion that subgroup differences disappear when stereotype threat is removed. If these were isolated incidents in the midst of extensive accurate depiction of Steele and Aronson's (1995) work, then these errors might not merit much attention. We thus sought to examine more systematically how Steele and Aronson's work has been characterized in the popular and scientific media. We conducted three examinations. The first looked at characterizations of Steele and Aronson in the popular media (i.e., magazines and newspapers). The second looked at characterizations of the work in scientific publications (i.e., journals and book chapters). The third looked at characterizations of the work in introductory psychology textbooks. In each case, we limited our examination to articles or textbook discussions that explicitly described the Steele and Aronson studies. Many more sources discussed stereotype threat more generally, without purporting to specifically present the findings of Steele and Aronson.

Note that in presenting Steele and Aronson's findings, an author can focus on within-group effects, between-groups effects, or both. We found that discussions of threat research that focused on within-group effects were not prone to misinterpretation. Such presentations compared African American student performance under threat and no-threat conditions and properly noted that the research clearly showed that the performance of African American students differs under the two conditions. Presentations of threat research that focused on between-groups effects (e.g., African American vs. White) were prone to misinterpretation: It is here that appropriate interpretation requires taking into account the fact that adjustments were made for existing SAT differences. Thus, our categorization of treatment of Steele and Aronson's findings is restricted to accounts of the research that discuss between-groups effects. Accounts that specifically noted the adjustments for SAT differences were classified as correct. Accounts of the research that ignored the SAT adjustment and reported that, absent threat, the scores of the African American and White groups were the same were classified as incorrect.

Popular Media

We conducted an electronic search for all references to stereotype threat and to Claude Steele. Many discussed stereotype threat generally; we located 16 articles that explicitly described Steele and Aronson's (1995) findings with regard to the relative performance of African American and White students. We characterized 14 of the 16 (87.5%) as incorrect, as they incorrectly asserted—in a variety of slightly different ways—that subgroup differences disappeared in the nonstereotype-threat condition. The appendix contains a sampling of quotations from these articles.

Scientific Journals

As with the popular media above, we conducted an electronic search of a variety of electronic databases, including PsycLIT, Social Science Index–Expanded, Expanded Academic Index, and the LexisNexis Academic Universe, using the keywords *stereotype threat* and *Claude Steele.* We found 11 articles and chapters that explicitly described Steele and Aronson's (1995) findings. We characterized 10 of the 11 (90.9%) as incorrect, as they incorrectly asserted that subgroup differences disappeared in the nonstereotype-threat condition. The appendix contains a sample of quotations from these sources.

Psychology Textbooks

We obtained a sample of 27 introductory psychology textbooks published since 1999 that had been sent to our department for course adoption consideration. We found that 18 of the 27 (67%) include a treatment of the topic of stereotype threat. Nine of the texts limited their discussion to within-group effects (e.g., stating correctly that African American students had higher test performance in the no-threat condition than in the threat condition). Nine texts made between-groups (e.g., African American–White) comparisons. Five of the 9 mischaracterized the findings by stating that the two groups performed equally in the no-threat condition. Thus, 56% of texts that discussed African American–White comparisons did so incorrectly. The appendix contains a sampling of quotations from these sources.

We can only speculate as to causes of the mischaracterization of the Steele and Aronson (1995) findings in these various media. One possibility is that authors of these articles and texts did not notice that test performance had been adjusted for prior SAT scores. We have anecdotal evidence to this effect, as in the course of our research on this topic, we have had numerous conversations with colleagues familiar with stereotype-threat research who expressed surprise when we informed them that adjustment had been made for prior SAT scores (including some who did not believe us until we produced the original article). A factor contributing to not noticing the adjustment may be the appeal of the misinterpreted findings (i.e., the conclusion that eliminating stereotype threat eliminates African American–White differences). Finding mechanisms to reduce or eliminate subgroup differences is an outcome that we believe would be virtually universally welcomed. Thus, research findings that can be interpreted as contributing to that outcome may be more readily accepted with less critical scrutiny.

A second possibility is that authors did not realize the implications of the fact that test scores had been adjusted for prior SAT scores. As an example, one psychology text (Passer & Smith, 2001) reproduced the figure from Steele and Aronson (1995) that we have included here as Figure 1A, but with one key exception: The parenthetical phrase "adjusted for SAT scores" has been eliminated from the *y*-axis. Thus, an active decision was made, either by the authors or by the textbook editorial staff, to remove the reference to adjustment, a decision that we believe would not be made if its implications were understood.

A third possibility is that the omission of reference to adjustment for prior SAT scores was an inadvertent error by authors who do recognize the implications of the adjustment. We offer as an example an article whose authors include the original researchers. Our appendix includes a quotation from Aronson et al. (1999) that discusses eliminating the African American–White gap without noting the adjustment for SAT scores. These authors have noted the adjustment in other depictions of their original work (e.g., Steele, 1997; Steele & Aronson, 1998).

Conclusion

We suspect that many readers may react with disappointment to our showing that the Steele and Aronson (1995) research does not show that eliminating stereotype threat eliminates African American–White test differences. Subgroup differences in performance on high-stakes tests represent one of American society's most pressing social problems, and mechanisms for reducing or eliminating differences are of enormous interest. Yet, given the importance of the problem, proposed explanatory mechanisms merit careful scrutiny and clear understanding.

Our concern about the misinterpretation that removing threat from a testing setting eliminates African American–White differences is that such misinterpretation has the potential to wrongly lead to the belief that there is less need for research and intervention aimed at a broad range of potential contributing factors, such as differences in educational and economic opportunities of African American and White youth. If group differences in scores on the SAT and other tests were largely explainable by the mind-set with which examinees approach the testing situation, it would then follow that differences in factors such as quality of instruction or per-pupil educational expenditure do not matter much in terms of achievement in the domains measured by high-stakes tests. Hence, caution in interpretation of threat research is warranted.

We wish to address several issues raised by readers of early drafts of this article. One is that misinterpretation of research is regrettably all too common and thus that documenting misinterpretations in one single domain is of limited interest. Our response is that we are singling out this domain because the issue at stake is of such importance and because the interpretive errors are so rampant and so systematic. A second issue raised by readers is that although we offer evidence of misinterpretation, we do not show evidence of any serious consequences of this (e.g., decision makers misled by misstatements of research findings). One response is that heading off future interpretive errors is justification enough for the article. A second response is that in our applied work with organizations using tests for personnel decisions, we have frequently encountered individuals responsible for decisions about test use who repeated the misinterpretations that we document here.

We reiterate that nothing we report here is intended as criticism of Steele and Aronson's (1995) original research or as a challenge to the concept of stereotype threat as an important phenomenon with relevance to testing settings. Steele and Aronson clearly showed that imposing and eliminating stereotype threat can, in laboratory settings, affect the test performance of both African American and White students, and other researchers have extended this to other groups (e.g., gender, age). This is important in that it highlights the fact that test scores can be influenced by factors other than the examinee's true level of skill and achievement. At one level, this is well known: The whole notion of standardized testing is based on controlling extraneous features of the testing environment. What is novel, though, is the demonstration that a standardized feature of test administration (e.g., the description of what the test measures) can have a different effect on one group of examinees than on another. Thus, continued attention to stereotype threat is certainly warranted. What we do here is caution against misinterpreting the findings as a complete explanation for the African American–White differences observed in Steele and Aronson's research and against prematurely generalizing the laboratory findings to high-stakes testing environments.

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.

Atkinson, R. L., Atkinson, R. C., Smith, E. E., Bem, D. J., & Nolen-Hoeksema, S. (2000). *Hilgard's introduction to psychology* (13th ed.). Fort Worth, TX: Harcourt College Publishers.

Bobko, P., Roth, P. L., & Potosky, D. (1999). Derivation and implications of a meta-analytic matrix incorporating cognitive ability, alternative predictors, and job performance. *Personnel Psychology, 52,* 561–590.

Chandler, M. (Writer & Director). (1999, October 4). Secrets of the SAT [Television series episode]. In M. Sullivan (Executive Producer), *Frontline*. Boston: WGBH.

Croizet, J. C., & Claire, T. (1998). Extending the concept of stereotype threat to social class: The intellectual underperformance of students from low socioeconomic backgrounds. *Personality and Social Psychology Bulletin, 24,* 588–594.

Cullen, M. J., Hardison, C. M., & Sackett, P. R. (in press). Using SAT–grade and ability–job performance relationships to test predictions derived from stereotype threat theory. *Journal of Applied Psychology*.

Davis, S. F., & Palladino, J. J. (2002). *Psychology* (3rd ed.). Upper Saddle River, NJ: Pearson Education.

Hartigan, J. A., & Wigdor, A. K. (Eds.). (1989). *Fairness in employment testing: Validity generalization, minority issues, and the General Aptitude Test Battery.* Washington, DC: National Academy Press.

Herrnstein, R. J., & Murray, C. (1994). *The bell curve: Intelligence and class structure in American life.* New York: Free Press.

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

Kosslyn, S. M., & Rosenberg, R. S. (2001). *Psychology: The brain, the person, the world.* Boston: Allyn & Bacon.

Leslie, C. (1995, November 6). You can't jump high if the bar is set low: A new prescription to help Black kids succeed. *Newsweek, 126*(19), p. 82.

McCarty, R. (2001, April). Negative stereotypes: A personal view. *Monitor on Psychology, 32*(4), 31.

Morse, J. C. (1999, December 27). Race and guts: Close to home. *Forbes, 164,* p. 165.

Oswald, D. L., & Harvey, R. D. (2001). Hostile environments, stereotype threat, and math performance among undergraduate women. *Current Psychology: Developmental, Learning, Personality, Social, 19,* 338–356.

Passer, M. W., & Smith, R. E. (2001). *Psychology: Frontiers and applications*. Boston: McGraw-Hill.

Passing the fairness test. (1999, October 5). *The Boston Globe,* p. A16.

Pulakos, E. D., & Schmitt, N. (1996). An evaluation of two strategies for reducing adverse impact and their effects on criterion-related validity. *Human Performance, 9,* 241–258.

Quinn, D., & Spencer, S. (1996, August). *Stereotype threat and the effect of test diagnosticity on women's math performance.* Paper presented at the 104th Annual Convention of the American Psychological Association, Toronto, Ontario, Canada.

Quinn, D., & Spencer, S. (2001). The interference of stereotype threat with women's generation of mathematical problem-solving strategies. *Journal of Social Issues, 57,* 55–71.

Sackett, P. R., & Ellingson, J. E. (1997). The effects of forming multi-predictor composites on group differences and adverse impact. *Personnel Psychology, 50,* 707–722.

Sackett, P. R., Schmitt, N., Ellingson, J. E., & Kabin, M. B. (2001). High-stakes testing in employment, credentialing, and higher education: Prospects in a post-affirmative-action world. *American Psychologist, 56,* 302–318.

Shih, M., Pittinsky, T. L., & Ambady, N. (1999). Stereotype susceptibility: Identity salience and shifts in quantitative performance. *Psychological Science, 10,* 80–83.

Steele, C. M. (1997). A threat in the air: How stereotypes shape intellectual identity and performance. *American Psychologist, 52,* 613–629.

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. (1998). Stereotype threat and the test performance of academically successful African Americans. In C. Jencks & M. Phillips (Eds.), *The Black–White test score gap* (pp. 401–430). Washington, DC: Brookings Institution Press.

Steele, C. M., Spencer, S. J., & Aronson, J. (2002). Contending with group image: The psychology of stereotype and social identity threat. In M. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 23, pp. 379–440). New York: Academic Press.

Stricker, L. J., & Ward, W. C. (in press). Stereotype threat, inquiring about test takers' ethnicity, and standardized test performance. *Journal of Applied Social Psychology*.

Wolfe, C. T., & Spencer, S. J. (1996). Stereotypes and prejudice: Their overt and subtle influence in the classroom. *American Behavioral Scientist, 40,* 176–185.

Appendix
Characterizations of Steele and Aronson's Findings

1. Examples From the Popular Press

"When students were told they were being tested for ability, the Black students performed more poorly than the White students. Was this because of stereotype threat? The researchers administered the test to other students, telling them the goal was to find out how people approach difficult problems. This time the researchers found no discernible difference between the performance of Black and White students." (Morse, December 27, 1999, in *Forbes,* p. 165)

"A Stanford psychology professor, Steele has done research indicating that Black students who think a test is unimportant match their White counterparts' scores. But if told a test measures intellect, Black students do worse than White students." ("Passing the Fairness Test," October 5, 1999, *The Boston Globe,* p. A16)

"In another experiment, when Blacks were told that they were taking a test that would evaluate their intellectual skills, they scored below Whites. Blacks who were told that the test was a laboratory problem-solving task that was not diagnostic of ability scored about the same as Whites." (Leslie, November 6, 1995, in *Newsweek,* p. 82)

2. Examples From Scientific Journals

"Steele and Aronson (1995) found, for example, that African-American college students were dramatically affected by stereotype threat conditions: they performed significantly worse than Whites on a standardized test when the test was presented as a diagnosis of their intellectual abilities, but about as well as Whites when the same test was presented as a nonevaluative problem solving task." (Aronson et al., 1999, p. 30)

"For example, Steele and Aronson (1995) found that when African American and White college students were given a difficult test of verbal ability presented as a diagnostic test of intellectual ability, African Americans performed more poorly on the tests than Whites. However, in another condition, when the exact same test was presented as simply a laboratory problem-solving exercise, African Americans performed equally as well as Whites on the test. One simple adjustment to the situation (changing the description of the test) eliminated the performance differences between Whites and African Americans." (Wolfe & Spencer, 1996, p. 180)

"Similar research found that African American participants' performance was impaired by making salient the stereotype that minorities perform poorly on diagnostic standardized tests (Steele & Aronson, 1995). African Americans performed equally to their White counterparts when the diagnostic use of the test was eliminated, thus eliminating stereotype threat." (Oswald & Harvey, 2001, p. 340)

3. Examples From Psychology Textbooks

"The results revealed that African-American students who thought they were simply solving problems performed as well as White students (who performed equally well in both situations). By contrast, the African-American students who had been told that the test measures their intellectual potential performed worse than all the other students." (Davis & Palladino, 2002, p. 358)

"When told that the test was simply a laboratory problem-solving task unrelated to ability, the Black students did just as well as the White students. But when told that the test was a test of intellectual ability, the Black students did less well than the White students." (Atkinson, Atkinson, Smith, Bem, & Nolen-Hoeksema, 2000, p. 615)

"African-Americans and Whites did equally well when told that the test was simply a laboratory experiment, but African-American students did much worse than Whites when they thought the test measured intelligence." (Kosslyn & Rosenberg, 2001, p. 284)

Comment

Contents

DOI: 10.1037/0003-066X.59.1.47

Stereotype Threat Does Not Live by Steele and Aronson (1995) Alone

Claude M. Steele
Stanford University

Joshua A. Aronson
New York University

Sackett, Hardison, and Cullen (2004, this issue) have raised a concern: that 29 mischaracterizations of an experiment from Steele and Aronson (1995) spread over eight years of media reports, journal articles, and textbooks could mislead teachers, students, researchers, policymakers, and parents into believing that the African American–White test-score gap is entirely caused by stereotype threat and not at all by group differences in opportunities and test-related knowledge, and that this belief could undermine efforts to improve African American students' academic skills. And, of course, the fact is that as much as we would like to (a) have a complete silver-bullet cure for the race gap and (b) have control over how our research is represented, we do not have either. So we too worry about mischaracterizations of research—an all-too-common problem. But we also worry that the Sackett et al. review overstates the threat in the present instance and may foster misunderstandings of its own.

The problem is this: In the universe of stereotype threat material that now includes over 100 articles and dissertations and more media mentions, Sackett et al. (2004) focus on the reporting of only a single experiment from the first published article on stereotype threat. This extremely narrow focus greatly exaggerates (a) that experiment's impact on people's understanding of stereotype threat and its role in the race gap, (b) the importance of particular aspects of the experiment, such as its use of covariance analysis, and (c) what its results say about the role of stereotype threat in real-life testing. We address these three issues in turn.

First, to know how important the mischaracterizations of Steele and Aronson's (1995) Study 2 are, one has to know whether they actually distorted people's understanding of stereotype threat's role in the race gap. Sackett et al.'s (2004) review, however, is too narrow to answer this question.

In the literature and reporting on stereotype threat, there are many discussions of its role in the race gap that do not describe Study 2 in Steele and Aronson (1995)—many more than 29. By excluding these from their review, Sackett et al. (2004) never assess how often these discussions make the overclaim they worry about—that stereotype threat is the sole cause of the race gap. The overwhelming majority of these discussions, we believe, get it right; they depict stereotype threat as one of multiple causes of this gap. Or, when they do give stereotype threat a special importance, they are not referring to the race gap in the general population but to the race gap in some subsample in which African Americans and Whites have been equated on other factors that might affect their test performance—either statistically or by selection, as when selective colleges select members of both groups for having high academic skills and, thus, similar educational backgrounds. To the extent that the groups have been equated on other causes of the race gap, it may not be overclaiming to emphasize stereotype threat as a principal cause of any remaining gap (see Massey, Charles, Lundy, & Fischer, 2003).

Finally, accurate discussions of stereotype threat's role might even be found in some of the articles in which they found mischaracterizations of the particular finding from Steele and Aronson (see Aronson et al., 1999).

In the literature on the race gap itself, Sackett et al. (2004) assessed neither the frequency with which stereotype threat is mentioned nor how often it was distorted when it was mentioned. Between the two of us, we have a respectable knowledge of this literature (e.g., Bowen & Bok, 1998; Jencks & Phillips, 1998; Massey et al., 2003). In these discussions where the focus is on explaining the race gap in large segments of the population, we have never seen a claim that stereotype threat is the gap's sole cause. Whenever stereotype threat is considered, it is placed, where it belongs, as one of multiple causes of the gap. No attentive reader would come away from this literature with a different view.

Steele and Aronson (1995) is one of the few stereotype threat studies that focused on African Americans. This may be why Sackett et al. (2004) chose to focus on it so exclusively. But it is eight years old. Its characterization of stereotype threat can be checked against dozens of subsequent published stereotype threat studies and discussions. And it is a drop in an ocean of information about the race gap.

Thus, although trying to correct mischaracterizations of research findings can be worthwhile, the importance of doing so depends on evidence that the mischaracterizations in question are affecting general understandings. Without such evidence, one has to assume that the mischaracterizations are unnoticed, that they are understood to be what they are, mischaracterizations, or that, in other ways, evolving literatures have self-correcting capacities. Sackett et al. (2004) offered no systematic evidence that this particular mischaracterization has had any effect—personal communications notwithstanding. Moreover, the narrow design of their review distorts the picture of how stereotype threat and its

Vol. 59, No. 1, 47–55

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 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