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

by Paul R. Sackett, Chaitra M. Hardison, and Michael J. Cullen

(2004) *University of Minnesota, Twin Cities Campus*

James J. Heckman, University of Chicago

Econ 350, Winter 2023

THE UNIVERSITY OF CHICAGO

- 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):

- 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 non-threat 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 non-threat condition, a finding supportive of the idea that the presence of stereotype threat inhibits minority group performance.
- This finding 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).

- What 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:

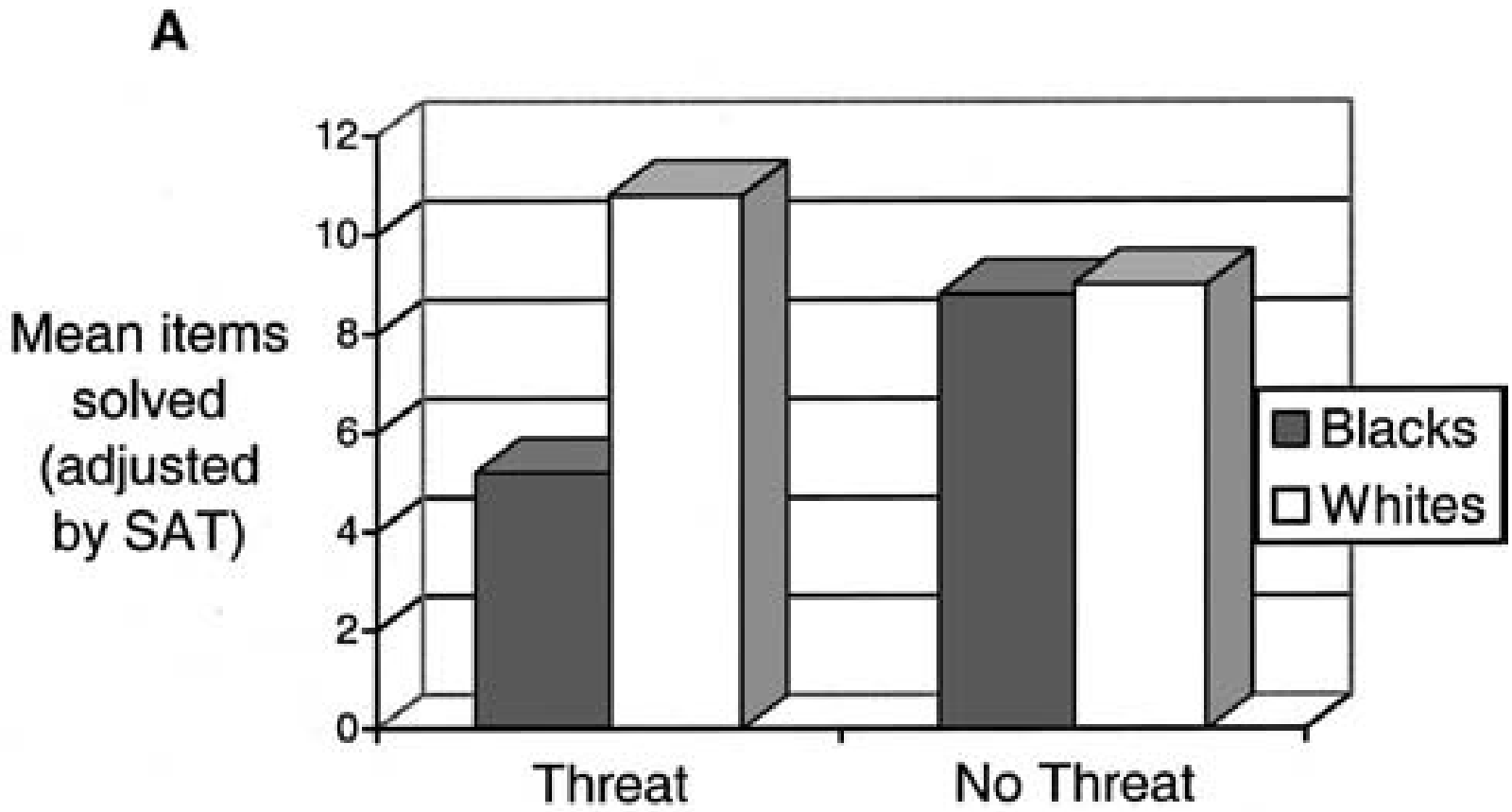
At Stanford University, psychology professor Claude Steele has spent several years investigating the 150-point score gap1 between Whites and Blacks on standardized tests. Was the cause class difference, lower incomes, poorer schools, or something else? In research conducted at Stanford, Steele administered 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.

- 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-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.

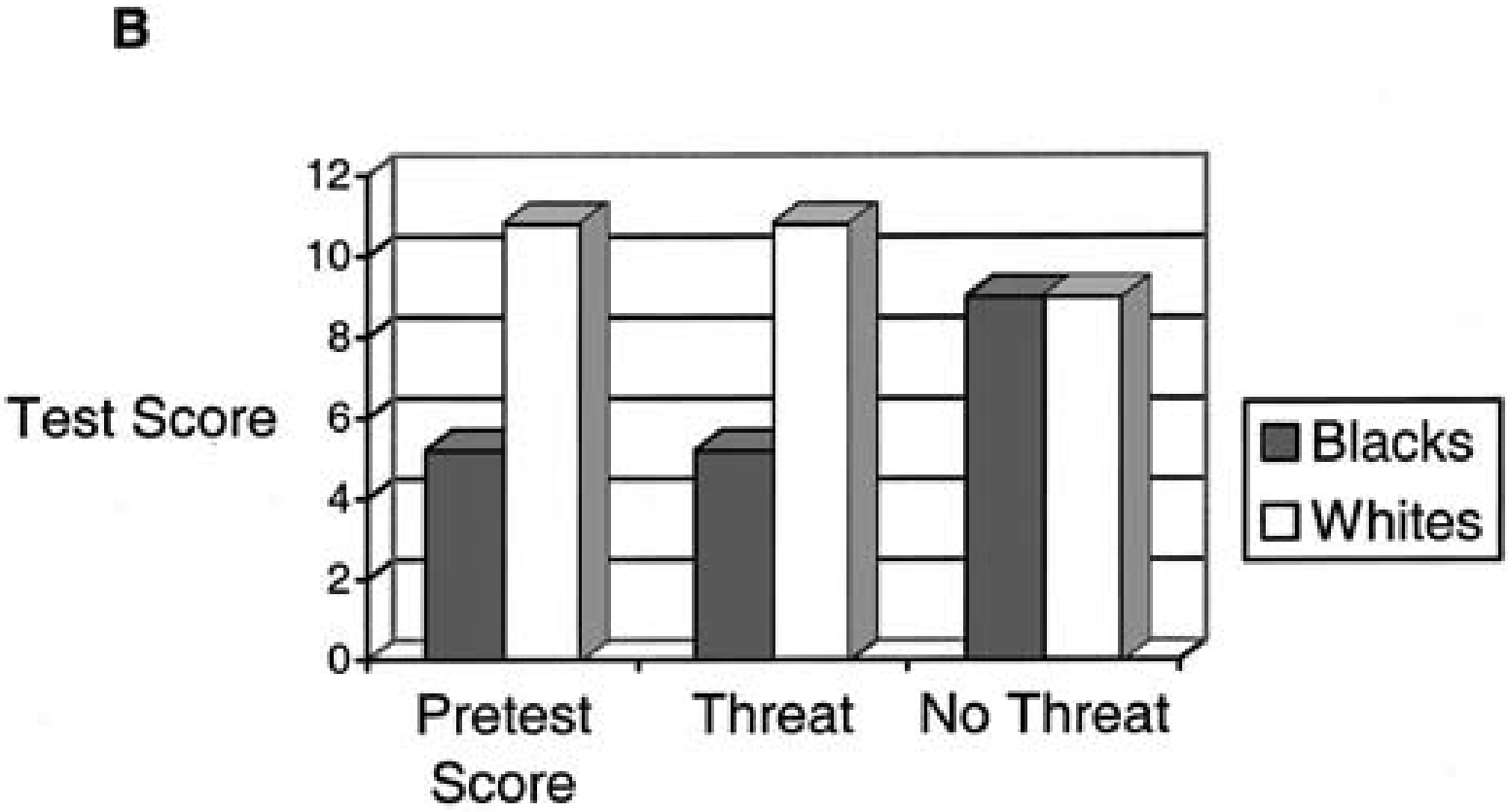
Figure 1: Interpretations of Steele and Aronson's Findings



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.

- 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.

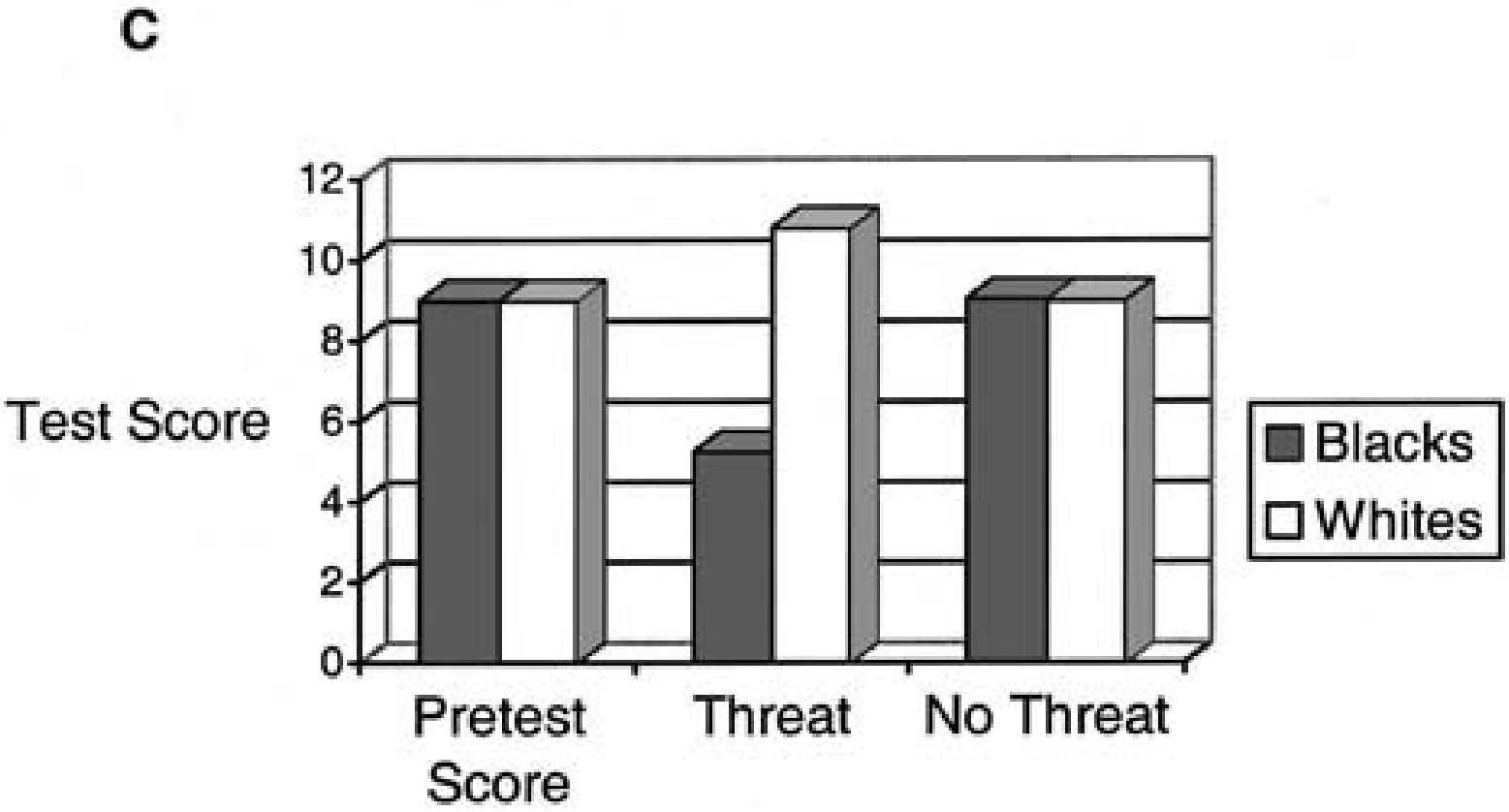
Figure 1: Interpretations of Steele and Aronson's Findings, Cont'd



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.

- 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.

Figure 1: Interpretations of Steele and Aronson's Findings, Cont'd



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.

- 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 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.
- Chetty (2018) makes exactly the same mistake.

Appendix

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

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 problemsolving 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)