Stereotype Threat in Applied Settings Re-Examined

Kelly Danaher and Christian S. Crandall[1]
University of Kansas

Stricker & Ward (2004) examined stereotype threat with a national sample of students taking an Advanced Placement (AP) Calculus exam, and a smaller sample taking Computerized Placement Tests (CPT). They inquired about gender either before (traditional) or after the test (which can reduce stereotype threat). They reported no significant effects of question timing. We reanalyze their findings, and argue that their conservative criterion for evidence led them to overlook significant stereotype threat effects with real practical implications. Women benefited substantially on the Calculus test, and on the CPT–Reading when demographics were asked after testing rather than before. This simple, small, and inexpensive change could increase U.S. women receiving AP Calculus AB credit by more than 4,700 every year.

In high-stakes standardized testing, men rule. As a group, men outperform women by a statistically significant margin on the SAT I (College Board, 2005), the Law School Admissions Test (LSAT; Dalessandro, Stilwell & Reese, 2005), the Medical College Admissions Test (MCAT; Association of American Medical Colleges, 2005), the Dental Admissions Test (DAT; American Dental Association, 2005), and the Graduate Record Examination (GRE; Educational Testing Service, 1999). With a few exceptions—notably Asian American men in math and natural sciences[2]—the highest scoring group in most high-stakes standardized testing is White men.

There is a very large debate about the sources of these differences among scientists, educators, policymakers, and the general public. This debate is fraught with politics, ideology, self-interest, and righteous indignation. Partisans, scientists, and the president of Harvard University have all come down on different sides of the issue. The key debate turns on whether the differences in test scores reflect differences in intrinsic aptitudes. Although nearly everyone agrees that different average scores reflect different levels of training and educational preparation, they differ on how much *other* factors play a role in determining scores.

One of the most intriguing ideas about the source of group differences on standardized test scores is the notion of stereotype threat. When a person is

[1]Correspondence concerning this article should be addressed to Chris Crandall, Department of Psychology, University of Kansas, 1415 Jayhawk Boulevard, Lawrence, KS 66045. E-mail: crandall@ku.edu

[2]On a few subscales, particularly the GRE–Quantitative, SAT Math, and some DAT–Science subscales, Asian American men have outperformed White men.

Journal of Applied Social Psychology, 2008, **38**, 6, pp. 1639–1655.

a member of a group that can be negatively stereotyped, there is an extra burden, a special risk: Their performance might be interpreted in terms of the stereotype. If the performance is consistent with the stereotype (e.g., a woman scoring poorly on a math test), it serves to confirm the stereotype to the audience. This is the phenomenon of *stereotype threat* (Spencer, Steele, & Quinn, 1999; Steele & Aronson, 1995). When the threat of confirming a negative stereotype is removed from a situation (e.g., by labeling a test as unimportant, by indicating that there is no sex or ethnic difference on a test), the performance of stereotyped group members often improves (e.g., Ambady, Paik, Steele, Owen-Smith, & Mitchell, 2004; Croizet & Claire, 1998; Marx & Goff, 2005; O'Brien & Crandall, 2003; Spencer et al., 1999; Steele & Aronson, 1995).

Most of the research supporting stereotype threat has come from laboratory settings, but there is also some evidence from field and classroom settings (e.g., Good, Aronson, & Inzlicht, 2003; Keller & Dauenheimer, 2003). As a result, although the main implication for stereotype threat is in classroom and high-stakes achievement testing settings (e.g., SAT, ACT, LSAT, GRE, MCAT), very little of the evidence for the phenomenon is linked to these tests in the field, as they are administered. It was with this problem in mind that Stricker and Ward (2004) conducted a series of studies that manipulated stereotype threat in the field. In this paper, we look closely at Stricker and Ward, and consider their findings' meaning for the practical applications of stereotype threat literature.

In what is probably the simplest version of stereotype threat experiments, Steele and Aronson (1995) simply manipulated whether Black and White participants identified their own racial category before answering GRE-type questions. In the "indicate race first" condition, White subjects outperformed Black subjects; but in the "race not indicated" condition, Blacks performed slightly better than did Whites (but not significantly so; see Steele & Aronson, 1995, Study 4). Implications for the testing procedure of standardized tests are simple and straightforward. As Stricker and Ward put it:

> Steele and Aronson's (1995) research on inquiring about ethnicity has obvious parallels with test administration procedures for widely used standardized tests that are employed in educational settings for admissions, course credit, course placement, and other purposes; and that require test takers to answer questions about their ethnicity and gender immediately before they take the tests. These parallels raise the real possibility that this practice may affect the performance of Black and female test takers on these tests. (p. 666)

Stricker and Ward (2004) examined the effects of stereotype threat on women and African Americans when taking the Advanced Placement (AP) Calculus Examination (Study 1) or Computerized Placement Tests (CPT) (Study 2). AP tests determine whether a student will receive college credit for college-level course work completed in high school, and CPTs determine future placement in college courses. These are important tests that determine future schooling. They are also interpreted as direct feedback about skill level. If stereotype threat lowers the performance of test takers, otherwise qualified students would receive unduly negative feedback about their abilities and would miss out on opportunities that passing scores on the tests provide, including courses completed, degrees earned, and career paths chosen. Research on stereotype threat has large implications for high-stakes testing.

In two studies, Stricker and Ward (2004), following Steele and Aronson (1995), altered the administration of the exams to manipulate stereotype threat, inquiring about ethnic and gender categories. In a set of two studies, Stricker and Ward were able to alter the administration of two real-life high-stakes tests, which had significant educational consequences to the test takers. Participants in these experiments were regular test takers, taking the tests under normal administration settings. In one study, high school students took the AP Calculus AB exam in their own high schools, and the results determined whether or not they received AP Calculus AB credit. In the other study, all incoming students at Central Piedmont Community College, in Charlotte, North Carolina took the CPT in the first few weeks as students at the college. In both cases, Stricker and Ward arranged for the administration of the test with either race and gender questions appearing before the main questions of the exam (this is typical administrative practice, which might increase the stereotype threat operating) or after (which should reduce, but not eliminate threat). In their paper, Sticker and Ward concluded that the manipulation had an effect that was so negligible that it might hardly matter at all:

> A clear and consistent finding in these two studies was the general absence of effects of inquiring about ethnicity and gender on performance on the two operational tests: the AP Calculus AB Examination, and the CPTs. No effects, negative or positive, that were both statistically and practically significant occurred, regardless of whether the students were Black, female, or from any other ethnic or gender group. (p. 685)

More recently, in response to a book review of a volume on gender and mathematics (Gallagher & Kaufinan, 2005), which appeared in *Science*

(Lewis, 2005), Stricker (2006) described the results of this study as follows: "In short, this study of actual test taking . . . found no evidence of the deleterious effects of stereotype threat on the performance of women on quantitative tests that have been observed in laboratory experiments" (p. 1310). Based on the data reported in their own article, we differ from Stricker and Ward's (2004) conclusion, and we do so vigorously. Stricker and Ward adopted what is a sensible strategy for data reporting. They accepted a result for consideration if it met two criteria designed to indicate both statistical and practical significance. They wrote:

> Both statistical and practical significance were considered in evaluating the results. For statistical significance, an .05 alpha level was used in all analyses (.05 was the familywise alpha level for the planned comparisons of simple effects, using the Bonferroni procedure, and for multiple comparisons with Tukey's test). For practical significance, a partial η (Cohen, 1973) of .10 in the ANOVAs and a *d* of .20 in the multiple comparisons were used. An η of .10 and a *d* of .20 represent Cohen's (1988) definition of a small effect size, accounting for 1% of the variance. (p. 674)

While we agree that this approach is sensible, especially given the very large samples that they collected, it represents a choice of favoring the acceptance of Type II error over Type I error. Simply put, the relatively high standards they make for accepting a result has two effects: It decreases the probability of mistakenly "finding" a null effect, and it increases the probability of missing a real effect. Like many choices in statistical analysis, the authors made a value judgment; that accidentally concluding that something is there (when it is not) would be worse than overlooking a real effect.

This is the basic value exchange that one always makes in setting a critical *p* value, and there is nothing inherently incorrect in what Stricker and Ward (2004) did. In fact, before we criticize them further, we wish to state that the authors were unusually complete, straightforward, and forthright in reporting their data. Their paper is a model of good presentation of results and design. It is only because of their full and complete reporting that we have the opportunity to reinterpret their results.

If one adopts a different set of values—and only slightly different—one would come to a very different conclusion about the data they report. Their choices have the effect of excluding what we think are important results from consideration; results very consistent with Steele and Aronson (1995), and with what we think are important practical implications. In this paper, we

adopt a very small difference in reporting standards, and we believe that the take-home message from the research changes plenty.

To emphasize by repeating, Stricker and Ward (2004) did not accept any effect unless it showed $p < .05$ in the overall ANOVA; $p < .05$ for planned comparisons (familywise, corrected by the very conservative Bonferroni method); and also showed $\eta \geq .10$ or $d \leq .20$. By these criteria, the manipulations of holding off inquiries about race or gender until after the test had no effect for the AP Calculus AB test or for any of the CPT scales.

However, if one uses the traditional $p < .05$ for the overall ANOVA (as did Stricker & Ward, 2004), and employs $\eta \geq .05$ as a standard, we find several important effects of the manipulation in Stricker and Ward's data. Steele and Aronson (1995) found that not inquiring about race improved the scores of Black students, but not Whites (an inquiry by stigmatized group interaction). Despite some minor differences of design, the comparable finding here would be either the Race × Inquiry interaction, or the Gender × Inquiry interaction. We searched for such interactions, considering the AP Calculus and CPT findings in turn, based on $p < .05$ and $\eta \geq .05$.

AP Calculus Study

The AP Calculus AB sample was based on nearly 2,000 participants from more than 70 different high school mathematics classrooms across the country. There were several selection criteria for inclusion in the sample (e.g., there must have been Black and White students in the class, teachers or administrators had to agree to the protocol). Stricker and Ward (2004) reported many different versions of the AP Calculus AB scores, including number attempted, number correct, accuracy rate, formula score, and AP grade. Although all of these variables are potentially interesting, we will focus on two dependent variables: formula score and AP grade. The *formula score* is basically the number correct, corrected for guessing, which provides the most basic overall continuous measure of performance. The AP grade is included because it is the single most practical measure: an *AP grade* of 3.0 or greater results in the student receiving college credit. The analysis that Stricker and Ward reported is a 2 (Timing) × 5 (Ethnicity) × 2 (Gender) ANOVA.

Stricker and Ward's (2004) data are consistent with the commonly found ethnicity and gender gap on standardized tests. They reported main effects on all dependent variables (number attempted, number correct, accuracy, formula score, free-response score, and AP grade), revealing gender and ethnicity differences, with girls significantly underperforming compared to boys, and Asian students outperforming White students, who in turn out-

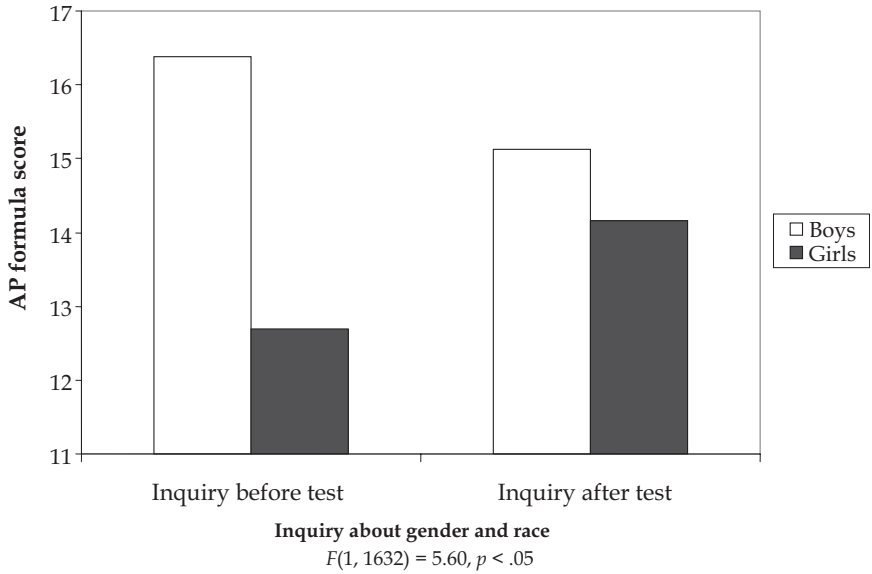

Figure 1. AP Calculus formula score by gender and timing of inquiry.

performed Black students (with "Others" and "Not Reporting" somewhere in between). Of more importance to our reanalysis, the relevant hypothesis test was the interaction between gender and timing, which was significant for both formula score, $F(1, 1632) = 5.60$, $p < .05$; and AP Grade, $F(1, 1632) = 4.16$, $p < .05$.[3] This interaction term was not significant for ethnicity.[4]

For gender, the interaction terms for the two most important variables were statistically significant at a conventional level. The results for formula score are presented in Figure 1, while the results for AP Grade appear in Figure 2. In both ways of calculating performance, the timing manipulation affected performance—women's AP Calculus AB grades and formula scores noticeably improved (and those of men declined) when participants were not asked about their gender beforehand.

[3]Of the six dependent variables, four had statistically significant Inquiry Condition × Gender interactions.

[4]The three-way interaction term was significant for formula score, $F(4, 1632) = 2.44$, $p < .05$. Based on the pattern of means (see Stricker & Ward, 2004, Table 3), the effect of the manipulation on gender for formula score was larger for Black girls than for White girls. This does not undermine the results presented in Figure 1, although it does suggest that the benefits of the manipulation would be greatest for Black girls. This three-way interaction term was not significant for AP grade.

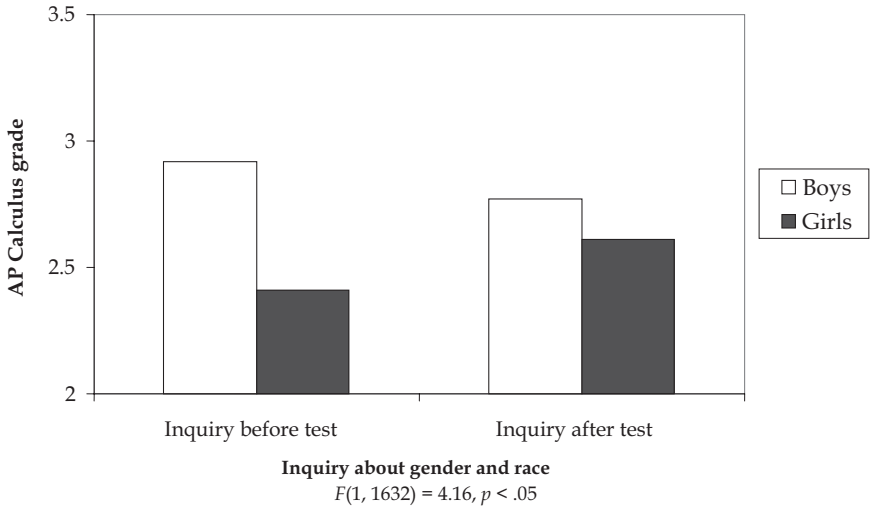

Figure 2. AP Calculus grade by gender and timing of inquiry.

Administering the test prior to the demographic inquiry diminished the gender difference in AP Calculus performance. How large is this effect? Is it practically important? Here, we calculate the practical size of the effect, based on the data from Stricker and Ward (2004). The cutoff for AP Calculus AB credit is a grade of 3, a number that is higher than any of the means reported. To illustrate the size of the effect, we have assumed that AP grade scores are normally distributed, with a standard deviation of 1.25 (from Stricker & Ward, 2004, p. 678). In Table 1, we present the means of the Gender × Inquiry condition in terms of standard-units distance from 3. By referring to the normal distribution, the standard-unit score means can be translated into the percentage of students who "passed" the test, which appears in the next column of Table 1.

The manipulation leads females to pass the test at about a 6% higher rate, males pass the test at about a 4% lower rate. This turns the difference between men and women in the passing rate from a whopping 16% down to a modest 5%; while the timing manipulation shrinks the sex difference by 33%. It is difficult to make the argument that this is not a change of large proportion, with real theoretical and practical significance. Put another way, under standard procedures, males perform at 149% of females when it comes to receiving AP Calculus AB credit. Under the modified procedure, this sex difference is reduced to 113%. Alternatively, girls achieved at 67% of the level of boys' performance under standard conditions, but improved to 88% of boys' performance under the modified timing.

Table 1

AP Calculus AB Performance by Gender and Inquiry Conditions

	Girls			Boys			
	AP grade	*Z* from cutoff	% receiving AP credit	AP grade	*Z* from cutoff	% receiving AP credit	Ratio of girls' success to boys' success
Inquiry before test	2.41	–0.47	31.9%	2.92	–0.06	47.6%	0.67
Inquiry after test	2.61	–0.31	37.8%	2.77	–0.18	42.9%	0.88

Note. AP = Advanced Placement examination. Z from cutoff = number of standard deviations the group's mean was from meeting the standard of 3.0 (i.e., point at which AP credit is earned); % receiving AP credit = percentage of test takers in that group that would qualify for AP credit.

Pragmatically speaking, the "trivial" differences carefully dismissed in Stricker and Ward (2004) can translate into very large practical effects, with real theoretical meaning. The inquiry manipulation reduces the gender difference to less than one third its original size. Instead of a ratio of about 6 girls receiving AP credit for every 9 boys who obtain credit, the new manipulation generates a ratio of about 8 girls receiving AP credit for every 9 boys.

How would this manipulation affect females at the population level of all students taking AP Calculus AB? Stricker and Ward (2004) told us that 52,465 boys and 47,275 girls took the test in 1995 (p. 669). Based on these numbers and the pattern of results illustrated in Figure 2, changing the way the tests are administered would increase the number of girls receiving AP Calculus credit from 15,081 to 17,870 in a year—an increase of 2,789 young women starting college each year with Calculus credit.

This size number should not be below the radar. The number of people taking the AP Calculus AB test is increasing. In 2004, there were 88,809 boys and 81,521 girls who took the exam (College Board, 2004), which represents an increase of 70.8% since Stricker and Ward collected data in 1995. All other things being equal, we estimate that 4,763 more women would receive AP Calculus AB credit if the timing were changed. We are convinced that stereotype threat in real-world testing situations can have a significant effect on test takers, and Stricker and Ward's (2004) data support this conclusion.

CPT Study

The CPT is a general test of knowledge and skills relevant to incoming freshmen. Stricker and Ward (2004) used a sample of more than 1,300 students that consisted of all incoming students for the fall term at a community college in Charlotte, North Carolina.

The CPT has four subtests: arithmetic, elementary algebra, sentence skills, and reading comprehension. Overall, the design of Stricker and Ward's (2004) study was the same as the AP study: Inquiry about demographics occurred either before or after taking the test. In this study, very few effects were found. Although main effects of ethnicity for all subtests and main effects of gender for elementary algebra and for arithmetic reflect the common test gap, the authors did not find gender differences on reading comprehension and on sentence-skills scores.

For one subscale (reading comprehension), Stricker and Ward (2004) reported the crucial interaction between inquiry condition and gender, $F(1, 1032) = 5.02$, $p < .05$. The pattern of means is displayed in Figure 3. These results appear very much like those in Figures 1 and 2, and represent quite typical stereotype threat results. Once again, women improved substantially

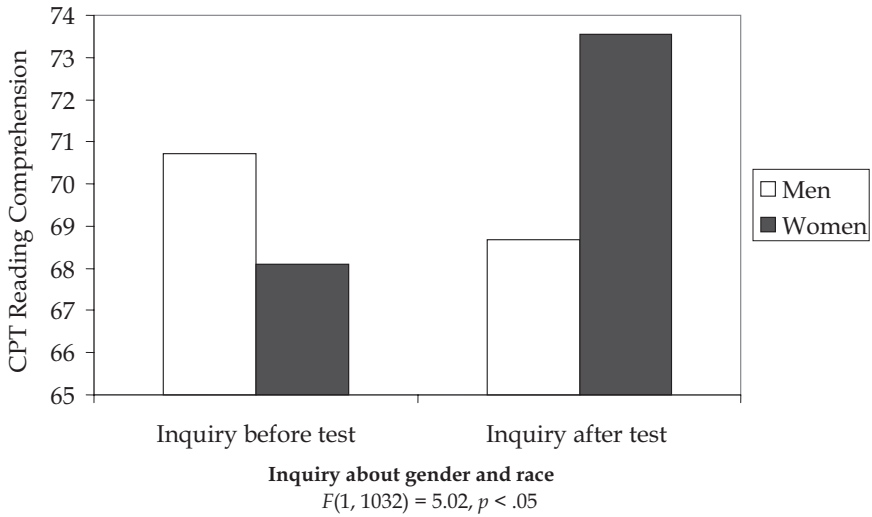

Figure 3. CPT reading comprehension by gender and timing of inquiry.

when they did not indicate their demographic categories, while men's scores declined slightly.

Because we do not have information about how CPT scores were treated as a cutoff, we cannot conduct the kinds of analyses that we reported previously for AP scores. However, we include these results to show that Studies 1 and 2 (Stricker & Ward, 2004) both demonstrate, to a significant extent, stereotype threat effects that replicate those of Steele and Aronson (1995).

Consequences of Analytic Strategies

With such practical implications, why did Sticker and Ward (2004) fail to report these findings in their results? First, Stricker and Ward set a conservative rule for finding effects (i.e., $p < .05$ *and* $\eta > .10$). But, there is nothing inherently good (or bad) about being statistically conservative. It is simply a preference for Type II error over Type I error. Stricker and Ward established η greater than .10 based on Cohen's (1988) definition of a small effect size, which they stated is not worthy of serious consideration. However, we feel that this recommendation and description constitute a (rare) major error on Cohen's part. One cannot label an effect as large or small without considering (a) the size of the effect in the context of other, similar effects; and (b) the practical importance of such an effect size. As

Stricker and Ward pointed out, it is true that an η of .10 represents an effect accounting for 1% of the variance. Such a "small" percentage accounted for is rather robust in real-world situations in which it is especially difficult to tease apart the many factors contributing to a given outcome (see Abelson, 1985, 1995; Prentice & Miller, 1992). Furthermore, an η of .08 in medicine or public health could add up to hundreds of thousands of saved lives or a Nobel Prize. Effects cannot be meaningfully judged solely by percentage of variance explained.

Stricker and Ward (2004) focused almost exclusively on simple effects testing (e.g., girls in the experimental group, compared to girls in the control group) within each ethnic and gender group as the most appropriate analysis for their data. They argued that a focus on simple effects was necessary, "given the specific hypotheses about Black and White students and females and males, and the need to compare the present findings with those of Steele and Aronson (1995)" (p. 674).

We do not find these reasons satisfying. First, both theory and practical considerations require that we compare stereotyped groups to non-negatively stereotyped groups. It is only in the comparison of the groups that gender differences have any meaning. In addition, because there is evidence of both stereotype threat (which harms the stereotyped) and stereotype lift (which privileges the non-stereotyped; Walton & Cohen, 2003), tests of the hypotheses must include both groups. The performance of females on tests matters under both conditions, but of the highest significance is how they perform in relation to their male counterparts. Removing ethnicity and gender inquiries prior to test taking diminished the differences in scores across gender, thus leveling the playing field for girls and boys. And here we agree with Stricker and Ward (2004), that "these analyses [the interactions between inquiry condition and ethnicity or gender] are informative in describing the actual effects of the experimental manipulations on the test performance of the AP and CPT test takers" (p. 689).

Stricker and Ward (2004) argued that between-group comparisons are inappropriate for their data because, unlike Steele and Aronson (1995), they were unable to control for prior ability. But the fact that the authors found significant interactions without partialing out other sources of variance demonstrates that the effect is powerful. Why would a direct comparison to Steele and Aronson's research be essential? There is no need to compare mean levels of performance; the testing situation and sample will almost always be different between different experiments, studies, or researchers. The critical issue here is testing experimental manipulations in situ; and on this critical test, stereotype threat effects emerged.

Sticker and Ward (2004) relied on the Bonferroni adjustment procedure, which is overly conservative in almost every application (Keselman, Cribbie,

& Holland, 1999; Moran, 2003), and most certainly conservative in a replication study in which the critical contrasts are known. Stricker and Ward were scrupulous at various points in their paper to hew to Steele and Aronson's (1995) original approach, replicating statistical tests whenever logically possible. Bonferroni adjustment typically is used for post hoc comparisons, and the statistical logic of the procedure presupposes accepting the null hypothesis as the foundation for calculation. Because the interaction F test was significant, because the hypothesis was not post hoc, and because this is essentially a replication study (hence the hypothesis is clearly a priori), Bonferroni is not appropriate. It is needlessly conservative, and it strongly privileges Type I error over Type II error. This is a value judgment that we cannot support in this context.

With their conservative rule for significance and size, with its consequence for analyses, Stricker and Ward (2004) concluded that "a clear and consistent finding in these two studies was the general absence of effects of inquiring about ethnicity and gender on performance on the two operational tests: the AP Calculus AB examination, and the CPTs" (p. 685). We believe that given Figures 1, 2, and 3 in the present paper; the significant hypothesis tests; and the analyses of population-level effect sizes, Stricker and Ward are inaccurate in their conclusions.

Perusal of the results reported by Stricker and Ward (2004) and our Figures 1, 2, and 3 show that girls perform better (5.9% more passing AP grades) when they are not asked about their category memberships. It also shows the *mirror-image effect*; that is, boys do worse without indicating their memberships (4.7% fewer passing AP grades). In every case here, and mimicking the rest of the literature (Walton & Cohen, 2003), the reduction of performance for boys is smaller than the increase in performance for girls.

Using the same calculation strategy for boys that we used for girls, we found that changing the timing of inquiry would have increased the number of girls passing by 4,763 in 2004, and would have decreased the number of boys passing by 4,211. Although slightly more boys take the AP Calculus test than do girls (from Stricker & Ward, 2004, we found that the percentages of boys taking the test were 52.4% and 52.1% in 1995 and 2004, respectively; College Board, 2004), because the reduction for boys was smaller than the increase for girls, the overall net benefit of combining boys and girls would be an increase of 552 students passing the AP Calculus test (based on the 2004 numbers). At the overall population level, the timing change would be very small. The formula score would improve from 14.63 under the current testing regimen to 14.67 with the postponed inquiry, and the grade score would improve from 2.678 to 2.694. These are very small differences.

At the individual level, these effects are not large, and this smallness is amply captured by Stricker and Ward's (2004) account. For AP Calculus

scores, it is of significant value only to those girls who are otherwise very close to passing, where the modest difference could be meaningful. (Likewise, the change would harm those boys who passed by a very small margin.) But at the population level, the effect might rightly be considered profound.

Stricker and Ward (2004) wrote that their research "assesses the generalizability of the laboratory findings to real life and evaluates the practical consequences of routine inquiries about ethnicity and gender in standardized test" (p. 668). We feel that their experiments have done this quite well. Our interpretation of their data suggests that putting off the assessment of ethnicity and gender until after the test itself has substantial effects. Stricker and Ward replicated Steele and Aronson (1995) in substantial detail, in a pragmatic field situation in which the stakes were notably high, the participants were highly motivated, and the consequences for future education were substantial.

Steele and Aronson (1995) found their timing effect comparing Black and White undergraduates, but Stricker and Ward (2004) found their effects with women. It is almost certain that different kinds of contextual changes in testing will have different effects for different stereotype-burdened groups. We do not have enough information from Sticker and Ward to ascertain what the "active ingredients" were. It may well be that gender stereotype effects might be slightly easier to erase in this kind of high-stakes testing, or it may be that Steele and Aronson's manipulation was more powerful (or more subtle), which is slightly more appropriate to ethnic stereotypes. It is too soon to tell, based on these data.

These results become doubly remarkable in the field context, for two reasons. First, Stricker and Ward (2004) chose what is probably the weakest of all successful stereotype threat manipulations: timing of demographic inquiry. This manipulation is so small that most test takers will probably not notice anything unusual about the testing situation. Second, it is hardly imaginable that changing the timing of inquiry removes stereotype threat from the situation. After all, this is an unmistakably important situation in which college credit and academic self-esteem are on the line; stereotype threat is certain to be high. The manipulation can only attenuate the threat, and only slightly in this context. Under what are among the most difficult of tests for the stereotype threat hypothesis, we find significant and practical support for the idea.

Stricker and Ward (2004) made what we believe is a fundamental error by equating practical significance with effect size statistics. According to Stricker and Ward, "For practical significance, a partial η of .10 in the ANOVAs and a d of .20 in the multiple comparisons were used" (p. 674). Meaningful consequences of a manipulation cannot be reduced to an arbitrary cutoff point of an effect-size statistic. The practical meaning of research results must

be understood in context, compared to costs and benefits, with careful attention to both the sample size of the research and the size of the population likely to be affected. Nearly 9,000 test takers would likely be affected by a timing change.

Stricker and Ward's (2004) research is especially valuable because it provides data that can be interpreted directly. Steele and Aronson (1995, Study 4) analyzed their participants' math performance after partialing out their SAT scores. Because of the possibility that Steele and Aronson's Black and White participants had different SAT scores, the two group means may have been adjusted differentially; thus, important group differences could not be studied directly (Sackett, Hardison, & Cullen, 2004). Sackett et al. pointed out that some consumers of this research have misunderstood the importance of this covariance-based analysis. Because Stricker and Ward did not adjust their dependent variables based on previous test scores, we can be doubly impressed at the reduction of gender differences in their data. Sackett et al. implied that stereotype threats may be limited to over- and underprediction effects. Stricker and Ward's data show that this is not the case.

We ask the reader to consider the imaginary Demeter Project (funded by the equally imaginary Drachma Foundation), which identifies girls likely to earn AP Calculus credit, and spends $11 per pupil per month over the course of 9 months, teaching calculus. This program is effective, in that it increases the rate of its students passing the AP Calculus exam from 30% to 45%. For the remarkably inexpensive yet effective Demeter Project to have the same size effect on the overall 2004 U.S. population that the inquiry switching does, the Drachma Foundation would have to reach 31,753 students, at a total cost of more than $3 million. The cost to testing bodies to change the timing of the demographic questionnaire surely would be much cheaper and would be a one-time expense. Changing the timing of demographic questions would be the single most cost-effective action our country could take to increase girls' performance on AP Calculus exams (cf. Lewis, 2005). Implementation would not erase gender or ethnicity differences, but it could occur almost immediately, the expense would be low, and the consequences for students at the beginning of their college careers could be quite high.

Summary

Let us summarize. Stricker and Ward (2004) replicated Steele and Aronson (1995) in a field setting, with very real consequences for performance. They found both stereotype threat and stereotype lift effects. These results were statistically significant, theoretically meaningful, and practically important. The results were strong and reliable in the AP Calculus study, and

somewhat supportive but not strong in the CPT study. Any review that suggests that the many laboratory studies showing stereotype threat and lift do not replicate in the field will be making the very same mistakes that Sticker and Ward (2004) made. We believe that Stricker (2006) was factually incorrect when he wrote

> This study of actual test taking, in common with our replication that examined community college students being given placement tests for algebra and arithmetic, found no evidence of the deleterious effects of stereotype threat on the performance of women on quantitative tests that have been observed in laboratory experiments. (p. 1311)

Let us end our discussion of Stricker and Ward's (2004) data where we began, with the question of values. Stricker and Ward made a series of value judgments when analyzing their data—a set of judgments that favor a conservative interpretation that is consistent with maintaining the status quo. They preferred to accept Type II error over accepting Type I error, which makes it more difficult to discover new things (see also Crandall & Schaller, 2002; 2004).

Although Stricker and Ward (2004) made their data and analytic strategy very explicit, the consequences of these choices were not as explicit. We must now make the question of values explicit with respect to our interpretation of their data. The experimental manipulation of inquiry timing has substantial effects, but are those effects good or bad? To answer this question, we must ask what we prefer: male success or female success at AP Calculus AB? Do we prefer large gender differences or small ones? Although we predict a very small increase in overall AP Calculus AB performance when inquiry comes post-test, the improvement of girls is largely matched by a decrease in boys' performance. Is there value to reducing a gender bias in math performance? Are girls underperforming relative to their "true" skills, and are boys overperforming? Are boys receiving a special privilege by indicating their gender or are girls suffering from an extra burden, or both? These value questions are difficult, but they deserve serious and explicit consideration, and they should not go unexamined.

References

Abelson, R. P. (1985). A variance explanation paradox: When a little is a lot. *Psychological Bulletin*, *97*, 128–132.

Abelson, R. P. (1995). *Statistics as principled argument*. Mahwah, NJ: Lawrence Erlbaum.

Ambady, N., Paik, S. K., Steele, J., Owen-Smith, A., & Mitchell, J. P. (2004). Deflecting negative self-relevant stereotype activation: The effects of individuation. *Journal of Experimental Social Psychology*, *40*, 401–408.

American Dental Association. (2005). *Dental Admission Testing program user's manual*. Chicago, IL: Author.

Association of American Medical Colleges. (2005). *Combined April/August 2005 MCAT performance*. Retrieved February 15, 2006, from www.aamc.org/students/mcat/examineedata/sum2005.pdf

Cohen, J. (1973). Eta-squared and partial eta-squared in fixed-factor ANOVA designs. *Educational and Psychological Measurement*, *33*, 107–112.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

College Board. (2004). *Advanced Placement program national summary report, 2004*. New York: Author.

College Board. (2005). *2005 College-bound seniors: Total group profile report*. New York: Author.

Crandall, C. S., & Schaller, M. (2002). Social psychology and the pragmatic conduct of science. *Theory and Psychology*, *11*, 479–488.

Crandall, C. S., & Schaller, M. (2004). Scientists and science: How individual goals shape collective norms. In M. Schaller & C. S. Crandall (Eds.) *The psychological foundations of culture* (pp. 200–223). Mahwah, NJ: Lawrence Erlbaum.

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.

Dalessandro, S. P., Stilwell, L. A., & Reese, L. M. (2005). *LSAT performance with regional, gender, and raciallethnic breakdowns: 1997–1998 through 2003–2004 testing years* (LSAT Technical Report 04-01). Newtown, PA: Law School Admission Council.

Educational Testing Service. (1999). *Trends and profiles: Statistics about GRE general test examinees by gender, age, and ethnicity*. Princeton, NJ: Author.

Gallagher, A. M., & Kaufinan, J. C. (Eds.). (2005). *Gender differences in mathematics*. Cambridge, UK: Cambridge University Press.

Good, C., Aronson, J., & Inzlicht, M. (2003). Improving adolescents' standardized test performance: An intervention to reduce the effects of stereotype threat. *Applied Developmental Psychology*, *24*, 645–662.

Keller, J., & Dauenheimer, D. (2003). Stereotype threat in the classroom: Dejection mediates the disrupting threat effect on women's math performance. *Personality and Social Psychology Bulletin*, *29*, 371–381.

Keselman, H. J., Cribbie, R., & Holland, B. (1999). The pairwise multiple comparison multiplicity problem: An alternative approach to familywise

and comparisonwise Type I error control. *Psychological Methods*, *4*, 58–69.

Lewis, D. (2005). Mathematics: Probing performance gaps. *Science*, *308*, 1871–1872.

Marx, D. M., & Goff, P. A. (2005). Clearing the air: The effect of experimenter race on target's test performance and subjective experience. *British Journal of Social Psychology*, *44*, 645–657.

Moran, M. D. (2003). Arguments for rejecting the sequential Bonferroni in ecological studies. *Oikos*, *100*, 403–405.

O'Brien, L. T., & Crandall, C. S. (2003). Stereotype threat and arousal: Effects on women's math performance. *Personality and Social Psychology Bulletin*, *29*, 782–789.

Prentice, D. A., & Miller, D. T. (1992). When small effects are impressive. *Psychological Bulletin*, *112*, 160–164.

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.

Spencer, S. J., Steele, C. M., & Quinn, D. M. (1999). Stereotype threat and women's math performance. *Journal of Experimental Social Psychology*, *35*, 4–28.

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.

Stricker, L. J. (2006). Stereotype threat: A clarification. *Science*, *312*, 1310–1312.

Stricker, L. J., & Ward, W. C. (2004). Stereotype threat, inquiring about test takers' ethnicity and gender, and standardized test performance. *Journal of Applied Social Psychology*, *34*, 665–693.

Walton, G. M., & Cohen. G. L. (2003). Stereotype lift. *Journal of Experimental Social Psychology*, *39*, 456–467.