1-1-2009

Stereotype Threat: a Case of Overclaim Syndrome?

Amy L. Wax
University of Pennsylvania Carey Law School

6

Stereotype Threat: A Case of Overclaim Syndrome?

Amy L. Wax

In math and science careers, men outperform women. Although many factors have been cited for these differences, the phenomenon of stereotype threat (ST) looms large as a favored explanation for observed disparities. Stereotype threat is a term coined by Claude Steele and his colleagues to refer to a psychological influence on test performance that derives from social expectations. The theory of ST predicts that, when widely accepted stereotypes allege a group's intellectual inferiority, fears of confirming these stereotypes cause individuals in the group to underperform relative to their true ability and knowledge. Men have long been assumed to possess superior talents in traditionally masculine fields such as mathematics and science. As a result, it is claimed, women face ST when attempting to perform in these domains.[1]

ST was initially described in a study investigating the reasons for the poorer performance of blacks than whites on standardized tests of academic aptitude. In an influential 1995 study authored by Claude Steele and Joshua Aronson, elite black and white Stanford University students were given an experimental test of verbal ability. Half were told the test would assess "individual verbal ability," while the rest were told that the purpose of the test was to evaluate psychological factors related to test performance. The authors theorized that the first instruction would call to mind stereotypes about blacks' inferior ability and thus would elicit an ST response, whereas the

I thank Jonathan Klick for helpful suggestions. Jason Levine and Alvin Dong provided excellent research assistance. All errors are mine.

second instruction would not have that effect. When resulting scores were adjusted for students' precollege scores on the verbal portion of the SAT (SAT-V), black students given the first (ST-diagnostic) instruction were found to perform below expectation, while those in the second (nondiagnostic) group performed as well as expected; white students, however, performed equally well under both conditions.[2] The authors concluded that the "threat"-induced fear of confirming stereotypes about black intellectual inferiority had caused black students in the threat-diagnostic condition to perform poorly.

In the wake of the Steele and Aronson paper, hundreds of studies and published journal articles have appeared that purport to document an impact for ST on test performance in a range of situations. Researchers claim that ST can depress test performance among lower socioeconomic classes,[3] Latinos,[4] the elderly,[5] and even groups that are not traditionally stereotyped.[6] Most notably, there is now a large body of work reporting that women perform worse on tests of mathematical skill under ST conditions—that is, when confronted with the stereotype of women's inferiority in math.[7] All in all, the phenomenon of ST has been analyzed extensively for over a decade and is now included in many standard psychology textbooks. Typing "stereotype threat" in a Google search yields thousands of relevant sites, many of which are mainstream media sources. ST has been repeatedly cited by newspapers, reported on television, and discussed in a variety of intellectual and political circles.[8]

It is not hard to see why advocates of social equality have seized on ST findings. If ST effects dominate, other causes of group performance disparities can be discounted. So, for instance, the Steele-Aronson observation that black students' verbal test scores are depressed under ST conditions suggests that longstanding test score disparities between blacks and whites might be due simply to performance anxiety rather than to real differences by race in academic ability, aptitude, or learning. The ST results also point decisively to broad social influences—most notably, invidious stereotypes and widespread assumptions of black inferiority—as the source of observed race gaps on commonly administered standardized tests, thereby banishing the bugbear of innate differences. But even conceding nurture, rather than nature, as the root cause of underachievement, attributing performance gaps to stereotype threat points away from arduous, long-term reforms like reducing discrimination or increasing a group's skill level. ST research raises the hope that underperformance is a short-term, situational problem that is amenable to the "quick fix"

of altering testing conditions or revising test instructions. The clear implication is that, if assumptions based on invidious stereotypes can be dispelled, the performance of lagging groups will dramatically improve, and test gaps will disappear.

The promise of an easy road to equality extends to gender. If women's situation-specific response to unjustified group generalizations is the source of observed gender gaps in scientific success, then other oft-cited factors—such as differences in ability, interests, drive, priorities, or temperament—can be discounted. ST research also promises a low-cost fix for women's underrepresentation in science. If the signals that cause women to achieve less can be dispelled, observed performance disparities will abate, and the accomplishments of men and women in scientific and quantitative fields will quickly equalize. In keeping with these observations, a psychologist writing in an American Psychological Association (APA) volume on women in science notes that

> the stereotype threat research carries two implications. First, if a simple manipulation of instructions can produce or eliminate gender difference in performance on a mathematics exam, the notion of fixed gender differences in math ability is called into serious question. Second, stereotype threat is a result of cultural factors—specifically gender stereotypes about female inferiority at mathematics—and thus provides evidence of socio-cultural influence on gender differences in mathematics performance.[9]

In the same vein, a report by the National Academy of Sciences, entitled *Beyond Bias and Barriers: Fulfilling the Potential of Women in Academic Science and Engineering*, regards ST results as confirming the conclusion that innate gender differences play essentially no role in observed patterns of scientific achievement and occupational success. Rather, states the report, gender differences are "strongly affected by cultural factors," which "can be eliminated by appropriate mitigation strategies, such as those used to reduce the effects of stereotype threat."[10]

This chapter is about whether ST explains observed differences in performance between men and women on standardized tests of quantitative skill, or in math and science careers more broadly. Is there reason to believe that ST is the sole, or even the primary, explanation for the underperformance

of females relative to males in these domains? After examining the key studies to date, the chapter concludes there is no basis as yet for identifying ST as an important, significant, or substantial contributor to observed gender disparities in test scores, academic achievement, or professional success in scientific fields. ST research to date has never shown that ST accounts for more than a trivial portion of observed gender gaps and thus fails to rule out a dominant role for other sources of female underperformance.

This is not to deny that the phenomenon of ST exists, nor that statistically significant ST-type effects have been demonstrated in some contexts. Many studies indicate that testing environment can interfere with test performance, with some groups perhaps more sensitive to these effects than others. Nonetheless, the ST literature raises serious questions about the significance of these results. The issue at the heart of ST research is this: How important is ST in explaining disparities in group achievement observed in the real world? More specifically, to what extent can gender differences in test performance and overall accomplishment be attributed to ST effects, as opposed to other causes? Does stereotype threat account for all, most, some, or only a little of women's underperformance relative to men on quantitative standardized tests and in scientific fields? Put more precisely, what percentage of the observed male–female gap in, say, math SAT (SAT-M) scores can be attributed to stereotype threat? In particular, what portion of the gap between men and women of outstanding ability—that is, those at the right tail of the bell curve who can be expected to comprise the great majority of high-achieving scientists—is due to ST?[11] These questions have not yet been squarely asked or answered. Despite the plethora of ST research, no study has precisely measured the magnitude of ST's effect relative to other influences on women's science and math performance overall. No study has told us "How much?" Yet that information will radically affect society's approach to women's underrepresentation in scientific fields. Specifically, if ST is the main culprit behind performance disparities between men and women, then resources should be directed almost exclusively to altering test instructions, improving women's working conditions, and countering the social stereotypes of women's lack of talent or interest in science. But if ST accounts for but a small portion of gender outcome differences, then efforts directed at manipulating testing conditions, boosting women's self-concept, or fighting social stereotypes are unlikely to yield significant results. Attention and resources are best expended

in other directions. Alternatively, if existing disparities express genuine differences in talents, life priorities, or preferences, gender gaps might prove relatively intractable to manipulation. The best strategy would then be to do little or nothing about gender disparities in science careers.

In addressing pivotal questions about ST's relative contribution to real-world patterns of gender performance, this chapter does not purport to take on all of the ST literature in detail. Nor is it meant to be an exhaustive, technical review of study results. Rather, it seeks to highlight certain patterns in the research that raise questions and concerns about its implications and the significance of the reported findings. For reasons already noted, the temptation to identify ST as the chief source of group performance differences is compelling. To borrow a phrase from another context, ST's powerful appeal gives rise to what has been dubbed "overclaim syndrome": the habit of ascribing greater weight to a body of scientific evidence than the data can bear.[12] It is, therefore, not surprising that, as Paul Sackett and his colleagues have shown, ST research has generated a number of sweeping and potentially misleading claims.[13] The goal of this chapter is to counter the temptation to overclaim syndrome as applied to gender by achieving a more balanced and measured view of the ST research results.

What are some of the problems with current research that leave open the question of how much ST contributes to observed gender disparities? First, there is the issue of relative magnitude: What is the size of the ST effect compared to the gender gap in performance overall, and to the gap observed in selective segments of the population? Second, what is the baseline yardstick for assessing ST effects? Do there exist reliable or objective measures of skill in math and science, impervious to ST, against which ST effects can be precisely gauged? Third is the question of the scope of ST's influence: Does ST operate as a "threat in the air?" Is it "out there" as a default condition, pervasively affecting women's performance in contexts routinely encountered in the real world? Is ST the ordinary and expected condition of test-taking—and, by extension, of doing science more generally—such that it can be assumed to undermine women's performance at all times and everywhere? Relatedly, does most research either support or assume that special interventions are needed to *dispel* ST (implying ST is pervasively "out there" in the background), or is it based on the premise that special interventions are required to *create* ST (implying ST is not ordinarily just "out there")? Fourth is the problem of

cherry-picking: Can the theory of ST explain why women do as well as or better than men in some measures of math performance (for example, grades in high school or college courses) but less well on standardized tests and in professional settings? And fifth, is there a novel approach to study design that might correct the deficits in ST research by generating crucial, missing information about the magnitude of ST's influence and its contribution to observed group differences?

The chapter concludes with a final challenge to ST research: if, as hypothesized, ST operates selectively to depress women's performance in math and science fields, how can that observation be reconciled with the full range of gender performance disparities, including those unrelated to quantitative domains? For example, why are women writers far less prominent and productive than men, even though women are widely believed to possess relevant talents that are equal to or better than men's? And what do these patterns imply for the plausibility of ascribing achievement disparities to ST more generally?

Relative Magnitude

Why do women's achievements in math and science fields fall short of men's? Because these fields draw heavily on quantitative ability, the attention of those seeking to explain these differences has been drawn to a longstanding gender gap in performance on the math portion of the SAT. The average scores of men and women on the SAT-M are not currently far apart, but the sex differential at the right tail of the bell curve, although fluctuating from year to year and narrowing somewhat over time, has always been substantial. In 2006, for example, the ratio of men to women scoring between 750 and 800 on the SAT-M was about 2.6 to 1.[14] This means that about 3.33 percent of the male test-takers scored in this interval, as compared to 1.29 percent of the females. The disparities are even greater in the upper reaches of this range. For example, between five and ten times as many boys as girls receive near-perfect scores on the SAT-M test in samples of mathematically gifted adolescents.[15] Student talent searches conducted at Johns Hopkins University yield similar ratios.[16] Since the most productive scientists are likely to come largely from this exclusive cohort,[17] it is important to investigate the sources of these

differentials. How much of this lopsided ratio on the SAT-M is due to stereotype threat? If ST were eliminated, would the ratio change much or at all? Would it disappear? Unfortunately, current research fails to answer these questions.

To understand why, it is necessary to take a closer look at actual ST studies. Most of the key research is performed on university students who are drawn from contrasting demographic groups (blacks and whites; male and females). The goal is to compare the test performance of students from each group under conditions designed to elicit stereotype threat and under circumstances that are not threat-inducing. Since it is only feasible to test each student once, subjects from each population must in turn be divided into an experimental category (tested under "threat" conditions) and a control category (tested under "non-threat" conditions), generating four separate subgroups overall. The goal is to conduct a four-way comparison, thus investigating if any difference can be shown in women's and men's performances under ST versus non-ST conditions.

Demonstrating an ST effect thus requires comparing test scores generated by four distinct groups of students. The problem is that the student subjects participating in any given study may have different levels of math ability. Accordingly, the average ability of students in each study category could differ as well. Thus, any observed difference in average test scores among the four groups of subjects in a particular study could reflect differences in ability rather than ST effects—or it could reflect some mixture of the two. And it is impossible to tell from the raw scores on an experimental test how much each factor contributes to observed patterns. For example, if the female "threat" subgroup scores worse than the female "control" subgroup, that could be because the study subjects in the first group are genuinely, on average, less able in math. Or it could be because the "threat" test condition depressed their scores. Likewise, if no such difference in performance is seen in men, that could be because men are not influenced by ST. But the same pattern of results would be observed if ST did, in fact, depress the performance of the male "threat" subgroup, but the men in that subgroup happened to possess greater average ability than the male "non-threat" controls. Taking the test under threat might then bring the average "threat" group score down to the average control-group level, creating the illusory impression that men are not vulnerable to threat.

To see this point, consider the following example. A researcher solicits student volunteers for an ST study. As is commonplace with these protocols, she chooses an equal number of male and female subjects, yielding twenty volunteers of each sex. She then randomly divides each group of twenty students into two groups of ten, to be assigned respectively to the experimental and control conditions. The men and women in the experimental group are given a math test under a "threat" condition. Those in the control group take the test under a non-threat condition. Assume that the average precollege SAT-M scores of students in each group turn out to be as follows:

- male non-threat (control)—590
- female non-threat (control)—590
- male threat (experimental)—605
- female threat (experimental)—550

The researcher then finds that women score significantly lower than men on the experimental test administered under "threat," but do as well as men when threat is removed. The results also show that men score somewhat higher under a threat condition than all other test groups. Assuming for purposes of this example that scores on the SAT-M reflect genuine math ability, does this observed pattern demonstrate an ST effect? The background SAT-M scores reveal that this pattern should not necessarily be interpreted this way. Rather, the scores on the experimental test might simply reflect average ability differences among the study subject groups. And even if test subjects are drawn from a relatively rarefied population—as would be the case for students attending a selective university—significant differences in ability levels could still exist.[18]

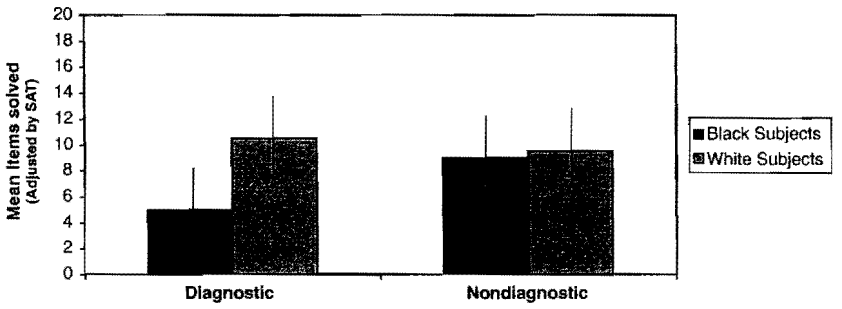
It follows that, in order to isolate and demonstrate any ST effects on test performance, subjects with similar background ability must be compared. There are two ways to accomplish this. The first is through statistical methods, such as adjusting performance for some reliable indicator of skill. Many researchers adjust experimental test results based on subjects' background SAT scores. This technique was used in the seminal 1995 paper by Steele and Aronson examining race differences in verbal ability,[19] and is employed in a number of gender studies as well. Alternatively, researchers use various

techniques to restrict test subjects' range of abilities more narrowly. This can be done more or less precisely. Some choose subjects who have obtained SAT scores within a particular interval. Others draw their study subjects from students enrolled in the same university course, or with the same course background, or with similar grades in particular courses. A number of gender studies take this tack.

All these methods omit key information critical to assessing the explanatory significance and policy implications of demonstrated ST effects. To see this, it is necessary to look more closely at actual research results. For their 1995 study of black and white Stanford undergraduates, Steele and Aronson solicited volunteers from the undergraduate population as a whole. They observed that, when their subjects' scores on an experimental verbal test were adjusted for the students' college entrance scores on the verbal portion of the SAT, the resulting adjusted scores were lower for blacks than whites under the designated threat condition (that is, when test-takers were expressly told the test would reflect verbal ability), but about the same when no threat was imposed. The authors interpreted the results as suggesting that, apart from any ability differences as reflected in SAT scores, ST conditions independently depress the test performance of black, but not white, students (see figure 6-1).

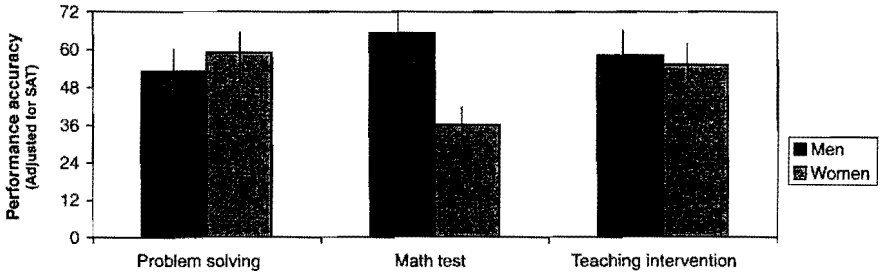
A similar method was adopted by Johns and collaborators in investigating ST's impact on women's math performance.[20] Their results showed that female students drawn from a college introductory statistics course performed worse than their male counterparts after hearing an experimental test described as "a math test" (which the researchers designated the diagnostic or threat condition), but just as well when expressly warned about the dangers of stereotype threat prior to taking the test (designated as the control or non-threat condition). The female test-takers also showed no shortfall in performance when informed that the test was designed to gauge general problem-solving skills (designated as a "teaching intervention," see figure 6-2). As with the Steele and Aronson study, scores on the experimental test were adjusted for each student's background SAT-M score so as to facilitate comparisons among the four distinct groups of subjects (male and female control, male and female experimental; see figure 6-2).[21] Once again, the study format was designed to isolate the effects of ST on test performance and to leave aside (by adjusting away) any performance differences among the subgroups that might be due to disparities in background ability.

FIGURE 6-1
STEREOTYPE THREAT AND THE INTELLECTUAL TEST PERFORMANCE OF AFRICAN AMERICANS



SOURCE: Steele, C. M. and J. Aronson. 1995. Stereotype threat and the intellectual test performance of African Americans, *Journal of Personality and Social Psychology* 69: 797–811.
NOTES: ST Condition (diagnostic instruction) = test problem solving ability; Non-ST Condition (non-diagnostic instruction) = determining psychological factors involved in solving verbal problems.

FIGURE 6-2
TEACHING STEREOTYPE THREAT AS A MEANS OF IMPROVING WOMEN'S MATH PERFORMANCE



SOURCE: Johns, M., T. Schmader, and A. Martens. 2005. Knowing is half the battle: Teaching stereotype threat as a means of improving women's math performance. *Psychological Science* 16, 175–79.
NOTES: Women's and men's accuracy scores (adjusted for quantitative SAT scores) on the math test as a function of the test description. Error bars represent standard errors.

Other gender studies, rather than controlling for SAT differences directly, seek to match ability level more or less precisely through criteria for selecting research subjects. For example, in evaluating women's math performance under threat and non-threat conditions, Spencer and colleagues tested twenty-eight men and twenty-eight women drawn from a psychology class at the University of Michigan.[22] In the first part of their study, the authors confined their sample to students who scored in the 85th percentile on the SAT-M (above 650). When subjects were tested under a "threat" condition—in which they were told that the experimental math test was one that revealed gender differences—the women performed significantly worse than the men. When the subjects were instructed that the test produced no gender differences, the performances of men and women were comparable. Although the reported data were not statistically adjusted for SATs and other ability-related parameters, the authors asserted that a data reanalysis using these adjustments did not alter the results.[23] This suggests that the subjects in their admittedly "highly selected" sample of research volunteers were roughly "equally qualified,"[24] and, thus, that underlying ability differences across their research subgroups (male, female, experimental, control) were probably insignificant.[25]

What are the implications of studies like these? As noted, to distinguish score differences due to ST effects from those reflecting disparate underlying ability, researchers must either choose subjects with similar ability or adjust their subjects' performance scores for some background measure of individual skill. Although these methods have the merit of helping to distinguish effects due to ST from those due to ability, they also create costs. First, as stressed by Sackett and colleagues, controlling for background ability or restricting the skill range of study subjects can potentially mislead by creating the unwarranted impression that stereotype threat is the exclusive source of group disparities in performance among the study subjects and, by extension, in the population as a whole.[26] This impression, although not justified by the research results, can arise from the way the results are presented. For example, graphs that display test scores adjusted for background SATs will often show little or no difference in performance between the relative comparison groups (such as black and white, or male and female, test-takers), despite the fact that the study subjects themselves—and the broader populations from which they are drawn—may differ significantly

in their ability levels as assessed by standardized test performance (see figure 6-1).

Likewise, studies that select subjects from a restricted range of ability levels can also create the misleading impression that *all* differences in test performance between groups (whether male–female or black–white) are due exclusively to stereotype threat. This impression arises from the fact that the range-restricted and ability-matched subjects in these studies are unlikely to represent an unbiased sample of the groups from which they are drawn. Because groups differ in their ability profiles, the degree to which a particular skill-restricted sample of subjects reflects the background population it represents will vary with each group. Indeed, in research designed to gauge ST effects by race or gender, study subjects matched for skill will almost certainly *not* be similarly representative of their background race or gender-specific population.

Consider a typical study designed to compare male and female math performance under stereotype threat conditions. Study subjects are chosen from students at a particular university. To qualify, all must have obtained a score of 750 or above on the math SAT. By definition, the men and women enrolled in the study will not be equally representative of the male and female populations as a whole. As noted, the ratio of men to women scoring above 750 on the SAT-M in 2006 was roughly 2.6 to 1. The male–female ratio toward the top of this range is even higher. Because women are significantly less likely to score above 750 than men, the female study subjects will be a more rarefied, and less typical, group than the men. In other words, the need to match the number and qualifications of study subjects across gender when investigating ST effects on women's math performance means that high-ability women in such research studies will be overrepresented, as compared to men, relative to their background same-sex population.[27]

The fact that men and women in typical ST studies are not likely to be similarly representative of their genders bears directly on whether these studies can answer the most critical questions: How big is the ST effect, and how much convergence in men's and women's scores can be expected from eliminating it? Consider once again the 2.6 to 1 ratio of male to female students scoring above 750 on the SAT-M. Would manipulations designed to dispel ST change that ratio significantly? Would altering test conditions elevate women's scores enough to match men's? The answer to that question

depends on what portion of the gender differential is due to ST. And that in turn depends on the magnitude of the ST effect relative to gender disparities in math ability that are unrelated to ST.

The problem is that ST's relative contribution to the observed gender gap cannot be calculated by using commonly employed protocols for looking solely at matched cohorts of students at the right tail of the bell curve—or, for that matter, at any restricted portion of the skill distribution. Because there are significantly fewer women than men obtaining the highest scores, many women lower down on the curve would have to improve their scores significantly to close the gender gap at the top. More precisely, the gender gap would not disappear unless women all along the distribution scored higher, shifting women up the bell curve until their numbers equaled those of men at each interval. It follows, however, that if gender disparities in standardized math test scores are due largely to ST test anxiety, those anxieties must be assumed to depress the scores of women at all levels of performance.

The question of whether ST depresses women's real-world test scores all along the curve by a sufficient amount to account for existing gender gaps cannot be answered by the current crop of ST studies. That is because those studies consider small numbers of subjects over a restricted range. Even if male and female subjects in a relatively small test sample are observed to do better—or equally well—on an experimental math test under specified non-threat, as opposed to threat, conditions, it cannot be inferred that changing the SAT to make test conditions less "threatening"—or manipulating standardized testing instructions for the population as a whole—will close or even significantly narrow the male–female gap in math SAT scores at any particular achievement level. The effects currently observed in a small, unrepresentative slice of women tell us nothing about whether anything like the necessary improvement in female scores overall would occur if ST effects could be reduced. It is just as likely that most of the gap in actual background test results is due to "real" disparities in math aptitude or problem-solving ability—disparities that will not yield to short-term manipulations but, rather, are the product of other types of long-term influences.

In sum, the protocols commonly used in ST research, which control for background SAT scores or draw study subjects from a narrow ability range, leave crucial information on the cutting-room floor. By deliberately abstracting away from overall group differentials due to factors other than ST, these

methods make it impossible to measure the magnitude of ST's contribution to score gaps relative to other causes. Because these studies provide no information about the comparative size of ST effects or the portion of existing background gaps that are due to ST, they tell us nothing about whether ST's influence is significant as compared to other factors like ability, knowledge, educational experience, interest in the subject matter, and learning.

For a concrete illustration of this problem, and of the potential for popular descriptions of ST research to mislead, consider a recent statement in a *New York Times* op-ed summarizing recent findings by Joshua Aronson and collaborators. The article states that "Mr. Aronson and others taught black and Hispanic junior high school students [that they] possessed the ability, if they worked hard, to make themselves smarter." According to the article, this intervention "erased up to half of the difference between minority and white achievement levels."[28]

The implication of this summary is that a large portion of the overall race achievement gap can be eliminated simply by telling students how capable they are. But this conclusion does not necessarily follow. We need to know far more about how this study was designed before leaping to such a dramatic conclusion. First, the op-ed report does not reveal whether the students in the study at issue were chosen randomly from the background population or whether they were matched for ability. Second, the description leaves us in the dark about the absolute magnitude of the ST-type effect observed relative to those subjects' test scores overall—or to any background achievement gap in the population as a whole. Indeed, we are told nothing at all about how the Aronson research was designed.

Consider one possible hypothetical scenario, which is fully consistent with the result reported in the op-ed piece. Suppose, as would be typical, that there is a significant disparity overall in the ratio of whites to minorities scoring in the top 10 percent on a standard junior high school achievement test. Suppose the ratio is 4 to 1, with whites even more dominant among the very top scorers. Suppose further that the subjects in the reported study were all selected to fall within that top 10 percent range. And assume, hypothetically again, that under high ST conditions, the white subjects in the Aronson study achieved an average score on the experimental test that was 5 percent higher than the minority subjects. Suppose also that under low ST conditions, that gap was reduced to a 2.5 percent average difference. Since 2.5 percent is half

of 5 percent, the experiment can thus accurately be described as demonstrating a testing intervention that "erased . . . half of the difference between minority and white achievement levels."

However, because the students in the hypothesized studies were matched for academic ability, such a 50 percent reduction of the score gap from a manipulation in testing conditions would not be surprising. ST could be expected to account for a relatively large portion of the residual group difference in performance among study subjects with similar abilities. Yet that result is consistent with ST having only a small *absolute* impact on the scores of the relatively able minority students in the sample. Although reducing ST cuts that impact in half, the reduction is against the base rate of a very small absolute effect. A 50 percent reduction in a small number is a small number. Thus, the reported 50 percent gap reduction could be entirely consistent with an ST effect that is quite small relative to the (otherwise similar) scores of the matched study subjects.

The more important point, though, is that a study that compares selected white and minority students of similar ability tells us nothing about the ability profiles of the groups from which they are drawn. Those profiles reveal large group differences—differences that are necessarily masked by any study protocol that matches subjects for ability. Moreover, the magnitude of the ST effects observed in such a study could well be negligible compared to the size of these group differences overall. Certainly it does not follow from the study results stated in the op-ed that eliminating ST can reduce this *overall* minority–white performance gap by 50 percent or anything close to that.

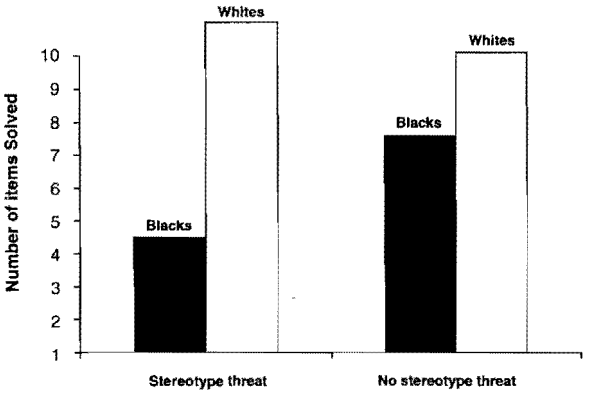
How do these insights apply in the gender context? Would redesigning studies of male and female test performance to report scores adjusted for SATs in conjunction with unadjusted or raw scores solve the problem? No. Although presenting data in this way has the potential to provide more information about the precise portion of the score gap in a particular study sample that is due to ST effects (as opposed to ability differences), it does not reveal the relative size of ST effects in the population as a whole. The problem is again one of representativeness. As noted, there is no guarantee that subjects of any study, or any subgroup in any study, are typical of the background population or even of a defined segment of that population. Likewise, it cannot be assumed—and indeed, given current study designs

and demographic realities, it is unlikely—that the comparison groups of subjects are equally representative, or typical, of their background populations.

This point applies to race as well as gender. In this vein, Steele and Aronson have recalculated their 1995 study results on ST effects on black and white students' verbal test performance using raw scores unadjusted for background SATs.[29] These numbers are summarized in an unpublished graph (see figure 6-3) provided by one of the authors.[30] When considered in conjunction with the SAT-adjusted data (see figure 6-1), the graph reveals that the black students in the study possessed lower average background verbal ability than the white students tested. It also shows that taking the test in the "threat" condition depressed black students' performance below the expected background levels, roughly doubling the preexisting racial performance gap.

Although examining both adjusted and unadjusted scores tells us something about the relative contribution of ST versus background skill to the black–white score gap in this particular study sample, it nonetheless fails to enlighten us on the contribution of ST to the black–white gap in SAT scores overall. Blacks scoring above 700 on the SAT-V are rare and much less common than white students scoring in this range.[31] Thus, the Stanford students tested for the Steele and Aronson study are not equally representative of blacks and whites as a whole, and may not even be similarly representative

FIGURE 6-3
VERBAL TEST PERFORMANCE UNCORRECTED BY SAT-V



SOURCE: Unpublished graph from Joshua Aronson.

of students of each race who score in these students' elite range. For this reason, the fact that being tested under "threat" significantly depressed the performance of a small group of black Stanford students tells us nothing about the extent to which test manipulations could alter the overall ratio of blacks to whites with superior scores on the SAT-V test. Nor does it tell us the degree to which reducing ST could cause SAT scores for blacks and whites as a whole to converge. As with gender, narrowing the gap would require an upward shift in scores all along the ability distribution. Given the magnitude of the existing black–white SAT gap, that shift would have to be dramatic indeed.

The Skill Baseline

The majority of ST studies reported in the literature compare test performance across distinct, non-overlapping groups of experimental subjects. The need presented by existing protocols to control for the skill level of study subjects poses the problem of how to measure real ability. Once again, an implicit assumption of many ST researchers is that "stereotype threat is an influence that may occur in an actual testing situation."[32] The implication is that the SAT-M gender gap—especially at the right tail—can be attributed mostly or exclusively to ST. But if ST does significantly depress SAT performance, then the practice of adjusting for or limiting the range of subjects' SAT scores begs the question of whether the SAT provides an accurate baseline measure of ability independent of the ST effect that the studies seek to assess. This observation points to a potentially fatal contradiction in the design of much ST research: SAT scores cannot simultaneously represent an accurate measure of math ability, untainted by ST, while at the same time being vulnerable to distortion by ST effects. If we accept that ST artificially depresses women's real-world test scores, then SATs do not reflect real math ability. Alternatively, if we posit that SATs are unaffected by ST, then ST effects cannot explain observed SAT gender gaps. Indeed, in that case, it is hard to see why we are interested in ST effects at all, since by hypothesis ST is irrelevant to the most important gender gap in real-world test scores!

In sum, ST researchers cannot have it both ways. They cannot use the SAT as an untainted, independent measure of ability and at the same time

claim that ST explains some, most, or all observed gender differences in standardized math-test performance. This inconsistency in the ST literature has been noted more than once in the context of both gender and race,[33] and has never been satisfactorily resolved. Rather, it has generated a number of confusing and contradictory statements. One research group, for example, has defended its use of an SAT control for comparing high-scoring male and female math students by observing that "performance-depressing stereotype threat emerged in these studies only when the test was at the limits of [students'] skills." The authors went on to conclude that "it is very unlikely that stereotype threat hampered [the women subjects'] performance on the SAT exam they had taken just a few years earlier. It too was well within their skills, as indicated by their high scores." They added, nonetheless, that, "over the full range," the performance of at least "some" women on the SAT-M was "likely" to be affected.[34] The problem with this explanation is that the SAT gender gap is largest in the highest score range.[35] These authors are therefore suggesting that where score disparities are greatest, ST is least likely to explain them. The clear implication of this suggestion is that the SAT-M score gap at the right tail is not due to ST—but rather to real gender differences in math ability, whether innate or acquired.

In another paper, however, scientists from the same group imply that the women of highest ability are most vulnerable to ST effects,[36] while women who "dissociate themselves from math at an early age," and thus get lower scores on standardized tests, are least likely to respond to ST.[37] In short, the literature is rife with waffling on a number of critical issues, including whether commonplace tests of math ability are tainted by ST effects at all, whether ST is responsible for differential performance only in selected portions of the ability distribution, and which women at which skill level are most affected.[38]

It should be noted that an important piece of evidence appears to undermine the assertion that ST systematically distorts women's real-world performance on the SAT-M—and thus supports the position that the test is an untainted measure of baseline math ability. The hypothesis that ST is largely responsible for the SAT-M gender gap generates a particular prediction about test results. If ST artificially depresses women's background SAT scores, then men and women with matching SATs should not perform equally well under experimental conditions that eliminate stereotype threat. Rather, women

should outperform men. Yet this pattern has not generally been observed.[39] These results cannot be squared with the claim that ST is an important source of group differentials in standardized test performance.

The Scope of ST's Influence

The issue of whether ST actually depresses performance on real-world tests is pertinent to yet another aspect of ST research, which is how ST experiments should be conducted. If real-world standardized tests are administered under conditions that are threatening to disfavored groups—so that observed score differentials can be largely attributed to ST—then it follows that ST is routinely present in ordinary testing situations. This means that ST is hovering out there "in the air" whenever anyone takes a test, so that no special measures or interventions are required to impose it. What are the implications of this assumption for experimental design? Because there is no need to create "threat," the administration of a test in the absence of any special instructions—or any instructions whatsoever—should constitute the diagnostic, experimental "threat" condition. In that case, however, creating the control, or non-threat, condition would appear to call for *affirmative* intervention. That is, the standing threat needs to be affirmatively removed or dispelled. Special instructions would therefore be needed to administer a test *without* the influence of ST.

Do social scientists consistently design their studies in keeping with these assumptions? Or do they implicitly assume that ST is not a pervasive background condition of all standardized testing, but rather taints test performance only in special circumstances? How do they generally define, identify, or create the experimental and control situations in ST research? How do they generate a "threat" testing condition, as opposed to a situation in which testing is free from threat? Once again, confusion reigns. Researchers in the field have not adopted a uniform protocol nor taken a consistent approach. In particular, the range of experimental designs reveals no consensus on whether ST is just out there "in the air," pervasively distorting the results of all standardized testing, or whether it is a condition that experimenters must create through special interventions or testing instructions.

To see this, consider the various ways researchers have generated threat and non-threat conditions. In one group of studies, scientists actively intervene to create the threat, usually by giving a specific pretest instruction. In this vein, researchers have told subjects about to take an experimental test that race[40] or gender[41] differences in test scores are to be expected. Or they have exposed subjects to gender-stereotypic television commercials prior to administering the test.[42] For the "control" or non-threat condition in these studies, in contrast, test-takers are either told nothing,[43] are given some kind of nongendered instruction (such as that the test is a gauge of personal math ability),[44] or are exposed to a stimulus (for instance, television commercials) with gender-neutral content.[45] These studies are generally most consistent with the implicit assumption that threat is not ordinarily "in the air," operating in most real-life test-taking situations, but rather must be specially created.

In contrast, other studies have researchers giving subjects special instructions for the purpose of dispelling or removing the threat. Thus, as reported in one paper, subjects in the control, or non-threat, group were told that the experimental test produced no gender differences and was "gender fair," while the "threat" (diagnostic) group was told nothing at all about gender.[46] In another study, the goal was to investigate "whether reminding women of other women's achievements might *alleviate* women's mathematics stereotype threat."[47] Thus, women who were about to take a difficult math test were informed that women make better psychology study subjects than men, or were read profiles of accomplished professional women. The expectation was that this group's performance would be unaffected by ST—that is, these instructions were supposed to generate a non-threat or control condition. In contrast, the "threat" group—which was expected to and did achieve lower scores—was given a gender-neutral reading about successful corporations. In yet another study, college-age mentors encouraged seventh-grade female subjects "either to view intelligence as malleable" or to ascribe their academic difficulties "to the novelty of the educational setting."[48] These student subjects delivered a better test performance than other girls who were given no such instructions. Studies of this type are more consistent with the assumption that all tests are taken "under threat," regardless of testing instructions. It follows that ST will operate to depress vulnerable groups' real-world performance unless specific steps are taken to blunt or remove its influence.

In still other articles, researchers used specific test instructions both to create and to dispel threat. In one, for example, some subjects were told that women were expected to do worse on the experimental test, while others were told that men and women performed equally well.[49] Yet other studies adopted a range of manipulations for comparing performance under supposed ST and non-ST conditions, including administering a test in mixed-sex or single-sex groups,[50] telling some women the test was designed to expose intellectual strengths while informing others that it highlighted intellectual weaknesses,[51] testing subjects in the threat and non-threat conditions in the presence of background noise while instructing some that the noise would likely depress their scores (that is, giving a so-called misattribution instruction in conjunction with an ST or gender-neutral condition),[52] and coaxing women into thinking more generally about their strengths rather than their stereotypical weaknesses.[53]

The dizzying array of research protocols raises obvious questions about the assumptions that inform these study designs. Specifically, when, if ever, must ST effects be affirmatively generated, and when must they be dispelled? What is the theory behind the answers to these questions, and what is the implication for whether and when ST operates on real-world testing? Can the so-called "threat" conditions in ST studies be analogized with real-life testing conditions? Are the study protocols consistent with the assumption that most testing—including math SAT testing—is conducted under "threat," or do they assume that most testing is free from threat? In other words, is there sometimes, often, or always a residual background ST effect "in the air"? Does threat require a special intervention—say, in the form of a gender-salient test instruction—or is it just "there" as the normal condition under which tests are generally taken, so as to require no special instruction? Why do some experiments show women performing as well as similarly skilled men when they are given no instruction (but underperforming after a threat-enhancing instruction), whereas others show women performing worse with no instruction (but performing just as well with a threat-dispelling instruction)? Is there an inconsistency here? One searches in vain for any analysis of these issues. Indeed, there is little systematic discussion in the ST literature of how theoretical expectations should inform research design, and virtually no consideration of whether the ST data as a whole are well-behaved in light of theory.

Cherry-Picking: The Selective Operation of ST

Yet further anomalies in the literature raise questions about the operation of stereotype threat within various domains that call upon math and science skills. Specifically, can ST explain the uneven pattern of female participation and achievement in these areas overall? Women are now about as likely as men to take advanced quantitative courses in high school and to major in math and science fields as undergraduates. That women earn better grades than men in high school and college math courses is often cited as evidence of their equal ability in these areas.[54] Yet women's enrollment in graduate school, their rates of professional advancement, and their productivity as working scientists lag behind.[55] Why does ST not diminish women's performance in the classroom or on class-related tests? Why are women not worried about confirming stereotypes in these contexts? The influence of ST would be expected here, especially in light of studies suggesting that mixed-sex settings (like coeducational college and university classes) generate ST threat effects and inhibit performance.[56] The few explanations offered—that, for example, standardized tests are generally intellectually demanding whereas coursework is uniformly "well within [women's] ability," or that women's experience of success within the classroom helps dispel stereotype threat[57]—are either questionable as a matter of fact (since upper-level math courses can be quite challenging) or circular (since women's record of classroom success just begs the question of why ST does not undermine that success in the first place). In short, attempts to account for observed patterns are, as yet, unsatisfactory.

Additional questions remain. Are ST effects cumulative and additive, or do they conform to an on–off pattern, such that someone either experiences the threat (with a fixed effect of determinate size), or not? If ST is "in the air," can researchers nonetheless further depress women's performance by giving a specific threat-generating instruction? Are ST effects on test performance linear in their impact—that is, do they sum up in a straightforward way? Different answers to these questions predict different results for ST research. The failure to match up theory to results—to come up with more precise hypotheses about how ST operates and then to devise studies designed specifically to confirm or disconfirm—is a serious flaw in the literature. These omissions represent yet another way in which social scientists have ignored important quantitative dimensions of ST.

Although ST research has so far been directed at validating the existence of the phenomenon, the next stage should undertake a more precise calibration of ST's magnitude relative to other influences on outcomes for men and women. The failure systematically and precisely to measure ST's impact over the full range of conditions makes it impossible to determine the size of ST effects as compared to other factors that can produce gender or group differences in performance. Yet knowledge of this relative magnitude is absolutely essential to an accurate assessment of ST's significance, which in turn is necessary to the development of a scientifically informed, rational strategy for dealing with differential group achievement. In particular, quantitative information is essential to any action plan for addressing gender gaps in math and science performance.

ST Study Design: Answering the Unanswered Questions

What questions should ST researchers now seek to answer? Put baldly, does ST account for 1 percent, 10 percent, 50 percent, 80 percent, or all of the gender difference in performance on standardized tests of math and science aptitude? What portion of the gender gap at the right tail of the bell curve—and in the number and achievements of the most productive scientists—can be attributed to ST? Addressing these questions requires measuring the background, real-world influence of ST, fixing a reliable baseline for its measurement, and gauging its relative contribution to existing disparities. These tasks cannot be accomplished without a paradigm shift in ST research. In particular, determining *how much* ST contributes to observed gender disparities calls for a radical new approach to ST study design.

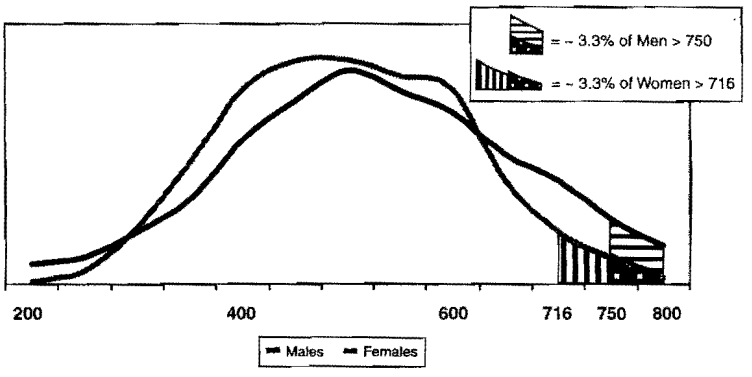
How might ST research be structured to reveal the pertinent information? More generally, is it possible to create a research protocol to address the key unanswered questions: Does ST account for all, some, or only a little of the gender gap in scores on standardized tests like the SAT-M (the relative magnitude problem)? Does ST significantly depress women's scores on standardized tests such as the SAT-M, or do such tests represent an accurate, untainted measure of real mathematical acumen (the baseline problem)?

One option is to begin with a well-defined working hypothesis. Although there is much equivocation on this point, assertions in the pertinent literature—

such as the statement quoted earlier from the APA volume addressing women's underrepresentation in the sciences[58]—strongly suggest that ST is a major, if not the sole, source of the gender gap in math and science performance. Therefore, one possible initial hypothesis is this: The gender gap in the SAT-M is due exclusively to ST. But if, in keeping with this hypothesis, it is assumed that "stereotype threat is responsible for the underperformance of women in quantitative domains," then it follows that "removing stereotype threat from those situations should eliminate women's performance deficit."[59]

How could this prediction be tested? That is, how could it be shown that eliminating ST's influence would close *all*—as opposed to some or none—of the gender gap in math and science performance? One possibility is to focus, as many gender studies already do, on a particular slice of the test-taking population—but to take a different approach. The women most likely to become prominent scientists are the ones at the extreme right tail of the bell curve—that is, women who score 750 and above on the SAT-M. As already noted, in 2006, 3.33 percent of male SAT test-takers scored between 750 and 800, while only 1.29 percent of female test-takers did so.[60] For purposes of illustration, a possible distribution of men's and women's scores consistent with these ratios is schematically depicted in figure 6-4.

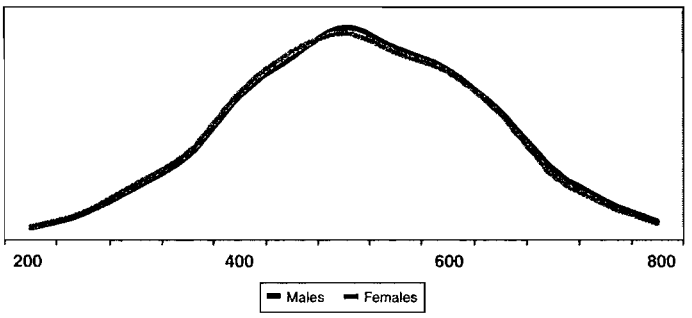
FIGURE 6-4
MATH SAT: "NORMAL" (STEREOTYPE THREAT) CONDITIONS



Source: Author's illustration.

Now, in keeping with our hypothesis, assume that men and women do "truly" possess equal math ability, and that the entire gender disparity for top scorers results from the operation of ST. A corollary of these assumptions is that, if stereotype threat could somehow be entirely dispelled, the percentage of women and men scoring 750 and above would precisely equalize. This means that the percentage of women test-takers scoring in this range would rise to 3.33 percent.[61] Accordingly, the distribution of men and women at the right tail of the bell curve would be the same. Indeed, the consequences of our hypothesis can be summarized more broadly: If the gender gap in SAT-M scores all along the distribution—including at the right tail—is due entirely to ST, then removing the influence of ST should cause the bell curves for male and female SAT-M performance to converge. That is, the percentage of males and females achieving each score would be equal. This result is schematically depicted in figure 6-5.

FIGURE 6-5
MATH SAT: STEREOTYPE THREAT REMOVED
(ASSUMING ST CAUSES THE GENDER GAP)



SOURCE: Author's illustration.
NOTE: Figure assumes ST causes the gender gap.

A comparison of the actual distribution of SAT-M scores (as reflected in figure 6-4) and the distribution (as reflected in figure 6-5) that would be predicted to result, in our hypothesis, from the removal of ST (if indeed ST is the sole cause of gender score disparities) makes it possible actually to measure the precise magnitude of ST's effect on women's SAT-M performance. The key is to

focus on the following question: Given the existing profile of SAT-M scores by gender, what is the score above which the percentage of women is equal to the percentage of men scoring 750 or higher? That is, what is the lowest score women would currently have to achieve to be in a group of equal relative size to that for men scoring at or above 750? Since 3.33 percent of men are in this range, we look for the minimum score actually achieved by the same percentage (3.33 percent) of women, which is roughly 716. Accepting our hypothesis, this allows us to estimate that ST depresses women's SAT-M scores, at least in this part of the ability distribution, by approximately thirty-four points.

This information is critical to determining whether our hypothesis is correct, because it permits us to decide whether ST in fact accounts for all, or some smaller part, of the gender disparity in SAT-M performance. Our hypothesis predicts that, if women scoring 716 or above on the math SAT could be retested without the influence of ST, their scores would significantly increase. More precisely, if ST is the sole cause of the gender gap, the scores of this cohort of women should rise to match men's—that is, to 750 and above.

How would we conduct this experiment? Ideally, it would be possible to identify women scoring above 716, and to select a cohort from this group that would reflect the distribution of women in this range; likewise for men scoring 750 or above.[62] Half of these men and women would then be asked to take an experimental math test under threat, and the other half in a non-threat condition. The performance of the men and women would then be compared. (The study's hypothesis is that the SATs are tainted by threat, which implies that threat is always out there "in the air." Consistent with this, the ST threat condition should involve administering the test with no special instruction, and the non-threat, or "control," condition would involve an instruction to dispel or eliminate the threat.)

What results would our hypothesis predict? In the threat—that is, normal testing—condition, the experimental test should show a gender gap that reflects the background gap in SAT scores for the study subjects. But administering the test under conditions that dissipate the threat should cause the gender gap to disappear. That is, the women in the study sample should achieve the same scores—on the same distribution—as the men. The bell curves in the subject groups should converge. In sum, if ST is the only reason for the observed SAT score gap, the male and female study subjects should, on average, achieve the same profile of scores in the non-threat

condition, despite women's lower background SAT scores. This reflects the understanding that, in the absence of threat, the same percentage of women as men will achieve each score.

Suppose that this result is not observed? It follows that our strong initial hypothesis—that ST is responsible for the entire SAT-M gender gap—is false. The experiment is nonetheless informative. Suppose, for example, that men's and women's scores narrow somewhat in the non-ST condition. Then measuring the extent of remaining divergence will allow a precise "decomposition" of factors responsible for the gender gap. Specifically, quantifying the remaining degree of divergence would enable researchers to measure exactly how much women's SAT-M scores are actually depressed by ST effects and how much of the gap is due to other influences. This would permit an assessment of the magnitude of ST's impact on women's SAT performance relative to other factors. This is the information that is currently missing—and just precisely what we are seeking.

The degree of gender-score convergence observed in this experiment also tells us something about the SAT as a baseline yardstick of "real" math ability. Indeed, if male and female scores in our experiment are observed to converge slightly or not at all, there is good news and bad. The good news is that the SATs look to be a true and objective measure of ability, unaffected by ST effects. Researchers would therefore be justified in adjusting experimental test results for background SATs as a way to compare subjects of unequal ability and to isolate the influence of ST. But the bad news is this: If the SATs are, indeed, a true and objective measure of ability untainted by ST effects, then it follows that ST can't be the source of the gender gap in SAT-M performance. That is, ST can't explain women's underperformance on these tests. But that begs the question of why we should care about ST effects at all. By definition, ST has little influence on the most important—and powerfully predictive—assessment of aptitude for math. It follows that the real reason for women's underperformance must lie elsewhere.

ST and the Problem of Pervasive Disparity

A final caveat on ST research is in order. In touting the influence of ST on women, social scientists have focused almost exclusively on performance in

selected areas—in particular, math and science. Because hoary stereotypes and traditional expectations about women's talents and interests have long held sway in these fields, the belief that pervasive cultural stereotypes impede women's performance in these arenas is widespread.

The problem with this selective focus is that women's underrepresentation in positions of achievement and influence is not confined to quantitative and scientific careers. Rather, men outperform women across the board, with women relatively scarce at the top of fields drawing on a broad range of aptitudes, including those for which women equal or outperform men on standardized tests and other well-accepted measures of ability. Dramatic gender disparities in achievement, productivity, output, occupational participation, and prominence persist even in areas where cultural beliefs regarding women's inferiority are absent, or where gender differences in ability have not been demonstrated, at least by conventional metrics.

Consider magazine writing, book authorship, and journalism. These endeavors require proficiency in writing and reading literacy—areas in which women are widely thought to excel and consistently outscore men on standardized tests.[63] Whether there are or ever will be equal numbers of men and women with the highest ability in math and science has been subject to vigorous debate, but few have suggested that women fall short of men in verbal skills. In light of these observations, the influence of gender stereotyping—and gender-based ST—is not generally believed to depress women's performance in these areas. Indeed, that women's achievement drawing on verbal abilities is unaffected by ST is an oft-stated assumption behind ST research designed to demonstrate the selective influence of ST on women's math and science performance.[64]

Yet women's "natural" verbal skills have not translated into dominance of fields drawing on these abilities. In particular, girls' strength in writing at all educational levels is not reflected in women's relative success in journalism or productivity in authorship of books and magazine articles. Among the books designated by the *New York Times* as the ten best of 2007, only two were written by women.[65] Of the thirty additional books recommended by the editors of the *New York Times* for 2007, seven were by women authors. In addition, the thirty-one winners of the 2008 Pulitzer Prize for writing and reporting included seven women.[66] Likewise, a routine perusal of advertisements by prominent publishing houses and university presses reveals a con-

sistent and pronounced predominance of male authors. A tally from recent publication lists confirms this impression. For books released by a sampling of scholarly publishers between January 2007 and March 2008 in history, philosophy, the social sciences (sociology, political science, psychology, economics, anthropology), public policy, and literature, men strongly outnumber women authors in all fields except literature.[67] Finally, an informal survey of pieces published in leading journals of opinion over the past three years reveals a decidedly lopsided pattern of authorship across the board, with male to female ratios of 28 to 1 for *Foreign Affairs*, 6 to 1 for the *New York Review of Books*, 7 to 1 for the *New Republic*, 6 to 1 for the *Atlantic Monthly*, and 4 to 1 for the *New Yorker*.

Can ST explain these dramatic disparities? Unlikely. But the persistence of wide gaps in productivity and achievement in areas conceded to be unaffected by ST casts doubt on ST's importance in math and science fields as well. Although the mix of factors leading to gender gaps need not be the same in all domains, the principle of Occam's razor suggests that those who would posit very different mechanisms for female underrepresentation across diverse fields bear the burden of persuasion. Given male dominance in occupations across the board—including many for which women are not stereotyped as less capable—it is important to step back and consider whether ST really accounts for most observed gender disparities. Factors that apply more broadly to many different endeavors should receive due consideration.

What unifying explanations can be offered? Perhaps authorship is not just a matter of verbal facility. Intellectual attributes of a more general kind, as measured by instruments such as IQ tests, may also be implicated. Although women and men are equal in average IQ, men outnumber women on the tails of the IQ distribution, with more men achieving the very highest scores.[68] It is far more likely, however, that women's relative lack of prominence is traceable to average gender differences in temperamental or "conative" traits such as competitiveness, ambition, singlemindedness, and drive,[69] or to women's greater attraction to and interest in people rather than things,[70] or to other gender disparities in patterns of intellectual interest,[71] focus on career advancement at the expense of domestic pursuits, or desire to achieve life balance.[72] In short, available evidence suggests that ST explains relatively little of the patterns of male and female accomplishment

observed in the real world today. Surely something else is going on. Although continued investigation of ST is certainly warranted, exaggerated claims for ST's significance should be avoided. A clear-eyed assessment of all the evidence is the only cure for overclaim syndrome.