RUNNING HEAD: Affirmative Meritocracy

Affirmative Meritocracy

Gregory M. Walton
Stanford University

Steven J. Spencer
University of Waterloo

Sam Erman
Harvard University

In Press at *Social Issues and Policy Review*

Author Note

We thank Charles Abernathy, Christopher Bryan, Geoffrey Cohen, Carol Dweck, Christine Logel, Richard Primus, Lee Ross, Claude Steele, Valerie Jones Taylor, Eric Uhlmann, Chris Whitman, and several anonymous reviewers for detailed and helpful input.

Abstract

We argue that in important circumstances meritocracy can be realized only through a specific form of affirmative action we call *affirmative meritocracy*. These circumstances arise because common measures of academic performance systematically underestimate the intellectual ability and potential of members of negatively stereotyped groups (e.g., non-Asian ethnic minorities, women in quantitative fields). This bias results not from the content of performance measures but from common contexts in which performance measures are assessed—from psychological threats like stereotype threat that are pervasive in academic settings, and which undermine the performance of people from negatively stereotyped groups. To overcome this bias, school and work settings should be changed to reduce stereotype threat. In such environments, admitting or hiring more members of devalued groups would promote meritocracy, diversity, and organizational performance. Evidence for this bias, its causes, magnitude, remedies, and implications for social policy and for law are discussed.

Keywords: academic achievement; affirmative action; latent ability; merit; stereotype threat; test bias

Affirmative Meritocracy

"[A]ffirmative action has to be made consistent with our highest ideals of personal responsibility and merit." - President Bill Clinton, National Archives, Washington DC, 1995

A fundamental problem in public life involves the perceived tension between diversity and meritocracy. When underrepresented ethnic minority groups or underrepresented gender groups perform worse than others on criteria used to make important admissions or hiring decisions, the value of creating diverse school and work settings and the value of admitting and hiring the most qualified candidates seem to collide. If a White student has an SAT-Math and -Verbal score of 1200 and a Black student has a score of 1150, how is it fair to admit the Black student over the White student? People tend to perceive hiring and admissions decisions based on applicants' potential to accomplish tasks as meritocratic and thus as fair and just (Bobocel, Son Hing, Davey, Stanley, & Zanna, 1998). Put differently, people advocate meritocracy, which we define as the systematic use of measures of potential to accomplish tasks in decision-making. If the White student is better positioned to perform well in college, should not this student be admitted? And yet if large ethnic group differences exist in a criterion like SAT scores, a selective university will admit few minority students. How to resolve this collision of values is the subject of much scholarly and public debate (e.g., Bowen & Bok, 1998; Crosby, Iyer, Clayton, & Downing, 2003; Herrnstein & Murray, 1994).

We argue that the perceived conflict between diversity and meritocracy is, in part, false. This conflict rests on a critical assumption: that measures of potential (i.e., merit) are fair and unbiased. Many people assume that grades and standardized test scores fill this role. Even if such measures do not predict subsequent intellectual or work performance perfectly, lay people and experts tend to assume that they are not systematically biased along important social dimensions

such as by race or gender (e.g., Jencks, 1998; Jensen, 1980; Sackett & Wilk, 1994). If this were the case, such measures would reward potential without discriminating unfairly or unjustly. But what if SAT scores, for instance, underestimate the ability and potential of disadvantaged students relative to advantaged students?

We review evidence that, as commonly assessed, common measures of "merit" are flawed in a way that perpetuates bias despite people's best intentions. Working in the tradition of stereotype threat (Inzlicht & Schmader, 2012; Steele, 2010; Steele, Spencer, & Aronson, 2002), Walton and Spencer (2009) recently found that common measures of intellectual ability systematically underestimate the academic potential of people from groups that are negatively stereotyped in intellectual settings. This bias results not from the content of performance measures but from psychological threats (e.g., stereotype threat) that are pervasive in common performance contexts like classrooms and testing situations. These threats depress the average test scores and grades of non-Asian ethnic minorities and of women in quantitative fields. It is as though some people run a race indoors while others run outdoors into a stiff headwind; thus some people can perform better even when they have no more ability or potential. Performance measures assessed in these settings underestimate the ability and potential of some people relative to others. This finding suggests that, despite a lower SAT score, a Black student who took the SAT under the burden of stereotype threat could be more qualified for college and likely to perform better in college than a White student.

If measures of merit are assessed in biased settings, hiring and admissions decisions based on them are not meritocratic. If group-based distortions infect such measures, only with remedies that mitigate this group bias can these measures become the basis of a meritocracy. In these cases, affirmative action and meritocracy entail some of the same principles of decision-

making.

This reasoning challenges common understandings of affirmative action, which typically assume unbiased assessments of merit. Many people oppose affirmative action because they believe that promoting diversity means degrading merit (Crosby, 2004; Crosby et al., 2003; Bobocel et al., 1998), a relationship presumed as well by the Supreme Court (e.g., *Rice v. Cayetano*, 2000, p. 517, where the Court argues that racial classifications cause "a person to be judged by ancestry instead of … merit"). But if measures of merit underestimate the potential of members of devalued groups, correcting this bias should promote both merit and diversity. Following past scholars (e.g., Bowen & Bok, 1998; Crosby, 2004; Crosby et al., 2003; Guinier & Strum 2001), we review the evidence that common settings cause systematic bias in common performance measures and discuss the affirmative steps organizations that value merit may need to take to admit, hire, pay, and promote people in unbiased ways.

We call this approach to affirmative action, "affirmative meritocracy." This approach extends past models by emphasizing changing organizational practice to combat stereotype threat (see also Logel, Walton, Peach, Spencer, & Zanna, in press). It involves two steps. First, organizations should create environments that minimize stereotype-related psychological threats. Doing so would mitigate one cause of achievement gaps and help create conditions in which all individuals can perform to their potential. Second, in making selection decisions, organizations should develop appropriate procedures to ameliorate the bias present in performance measures that systematically underestimate the ability and potential of members of devalued groups. These procedures should promote both merit and diversity.

By emphasizing bias that can inhere to the *contexts* in which performance measures are assessed, our approach complements research on biases that can inhere in the *content* of

performance measures. Several lines of research show that assessing a wider range of abilities than is typical can promote both merit and diversity. Sternberg and colleagues (2006), for instance, find that assessing college applicants' practical and creative skills, in addition to SAT scores and high school grades, can shrink group differences and raise correlations with college grades (see also Gardner, 1999). Similarly, Oswald, Schmitt, Kim, Ramsay, and Gillespie (2004) emphasize assessments of "noncognitive" qualities such as those related to leadership and interpersonal skills (see also Sedlacek, 2004). This research highlights the critical point that diverse qualities promote intellectual and professional success. Relatedly, biased decision-making results when an overly narrow predictor shows a larger group difference than exists on the criterion; relying on a narrow predictor can lead to the rejection of candidates with low scores on this measure but strengths on important unmeasured variables (i.e., *selective system bias*, Jencks, 1998). Other research highlights the importance of using broad measures to evaluate selection decisions (e.g., Guinier & Strum, 2001; Oswald et al., 2004). For instance, even when ethnic minority students perform worse in college than nonminority students, they may go on to careers and professional success that equal or exceed those of nonminorities (Bowen & Bok, 1998). We complement this past research by examining bias in traditional measures of academic ability—grades and test scores—and the origin of this bias in the psychological context of common academic environments.

This paper proceeds in three parts. First, we describe psychological threats that bias measures of intellectual performance, evidence for the bias, its magnitude, and its implications for the interpretation of group differences. Second, we discuss how organizations can reduce psychological threats and raise the performance of people from stereotyped groups. Third, we discuss implications for admissions and employment decisions.

Psychological Threats Bias Measures of Academic Performance

Measures of academic performance (e.g., test scores, grades) can be biased in several ways (Jencks, 1998). One way is that they can predict subsequent performance less strongly for one group than for another (bias in slope). A college admissions test could correlate less strongly with college grades for women than for men. Another way is that they can underestimate the subsequent level of performance of one group relative to that of another (bias in intercept). Women could earn better grades in college than men with the same score on the SAT. It has long been asserted that most modern assessments are not subject to significant bias of either type (Jencks, 1998; Jensen, 1980; Sackett & Wilk, 1994). But new data from several sources suggest that the second kind of bias is pervasive: Common measures of academic performance underestimate the ability of negatively stereotyped students relative to nonstereotyped students and do so throughout the performance range (Walton & Spencer, 2009).

Laboratory Research: Psychological Threats Undermine Academic Performance

This mean-level bias arises from the psychological context in which performance measures are typically assessed. When ethnic minority students perform in school, or when women perform in quantitative fields, they are often aware of stereotypes that impugn the ability of their ethnic or gender group. They may worry that a poor performance could lend credence to the stereotype. This worry, termed *stereotype threat*, can take up needed executive resources and undermine intellectual performance (Steele et al., 2002). In a classic series of studies, Black students performed worse than White students on a GRE test described as evaluative of verbal ability, an arena in which Blacks are negatively stereotyped. But when the same test was described as nonevaluative—rendering the stereotype irrelevant—Blacks performed as well as Whites (controlling for SAT scores, Steele & Aronson, 1995). In another classic study, women

told that men outscored women on a difficult math test performed worse than men. But when told that the test yielded no gender differences—refuting the stereotype—women and men performed equally well (Spencer, Steele, & Quinn, 1999). Stereotype threat can undermine performance in any group whose ability is negatively stereotyped in a domain, including among Latinos and people of lower socioeconomic status taking intellectual tests, the elderly taking memory tests, and even members of high-status groups when they face negative stereotypes (see Steele et al., 2002). But research demonstrates that performance decrements in school and work settings primarily disadvantage people who face pervasive negative intellectual stereotypes—generally, non-Asian ethnic minorities and, in math, science, and engineering, girls and women.[1]

Evidence that Psychological Threats Bias Real-World Measures of Academic Performance

Hundreds of laboratory experiments demonstrate the basic stereotype threat effect. But does stereotype threat undermine real-world academic performance (see also Aronson & Dee, 2012)? The hypothesis that it does is theoretically consistent—standardized tests, for instance, are typically taken in conditions similar to those that trigger stereotype threat in the laboratory; for example, they are represented as evaluative of stereotyped abilities. Nonetheless, the real-world effects of stereotype threat are controversial (Sackett, Hardison, & Cullen, 2004). As one critic charges, "No one has ever demonstrated that stereotype threat has any effect on black [students'] performance on actual tests that matter" (Sander, 2004a, p. 1996).

There are at least three ways to examine this question. First, if stereotype threat undermines real-world performance, then stereotyped students who experience more threat should do worse. In a sample of nearly 4,000 college students at 28 schools, Massey and Fischer (2005) found evidence for this hypothesis. Controlling for demographic variables and prior academic performance, Black and Latino students who reported experiencing more stereotype

threat subsequently earned lower GPAs. Importantly, this relationship held only for students taking classes with few ethnic minority professors, where stereotype threat is most relevant (see also Mendoza-Denton, Downey, Purdie, Davis, & Pietrzak, 2002). Another study found evidence that, when cues heighten stereotype threat, stereotyped students' performance differentially drops. Reardon, Atteberry, Arshan, and Kurlaender (2009) examined the performance of more than 60,000 California high school students on the state graduation exam, a test that is represented as evaluative of students' intellectual ability, as a function of performance on prior state tests, which are represented as evaluating schools not students and may thus induce lower levels of stereotype threat. Comparing students with the same level of prior performance, Reardon and colleagues found that Black and Latino students performed worse on the graduation exam than nonminority students and that, on the math portion, girls performed worse than boys. The effects held controlling for many factors, and the authors attribute them to stereotype threat.

These studies, however, are correlational and subject to alternative explanations. Skeptics (Sackett, Borneman, & Connelly, 2008; Sackett et al., 2004; Sander, 2004b; Wax, 2008) and testing organizations (e.g., the College Board, Kobrin, Sathy, & Shaw, 2007) have called for a second test. If stereotype threat undermines real-world performance, then measures of academic performance should underestimate the ability and potential of stereotyped students. In part, stereotyped students' ability would be *latent*—underestimated by their level of performance (Walton & Spencer, 2009). If this *latent ability hypothesis* is correct then, in nonthreatening environments, stereotyped students should perform better than nonstereotyped students with the same level of prior performance.

This is an important test. Until recently, it went unassessed. Although past survey studies compare performance measures, such as SAT scores and college grades (e.g., Cullen, Hardison,

& Sackett, 2004), they do not reduce threat on either measure. As such, these studies test only whether one measure is normally assessed in a more threatening environment and so is more biased than the other. But psychological threats may be pervasive in intellectual environments, undermining both test scores and grades (Steele, 1997). The critical question is whether measures of academic performance underestimate the performance of stereotyped students *in environments in which psychological threat has been reduced*. Stereotype threat laboratory experiments provide relevant data, but each has limited statistical power and is restricted to a particular population and testing instrument.

To provide a comprehensive test of the latent ability hypothesis, two of us recently conducted a meta-analysis of these laboratory experiments (Walton & Spencer, 2009). We tested whether, among students who had the same level of prior real-world performance, stereotyped students would perform better than nonstereotyped students in conditions that reduce stereotype threat. The meta-analysis included data from 39 laboratory experiments. Participants (n = 3,180) resided in 5 countries, ranged from elementary school students to college students, and included Blacks and Latinos in the United States, Turkish residents of Germany, and girls and women in both North America and Europe. Each experiment assessed students' prior real-world academic performance in the domain at hand (e.g., SAT score, grades), manipulated stereotype threat, and assessed students' performance on a test in the same domain. The studies used diverse strategies to reduce stereotype threat, including (1) refuting the stereotype (e.g., portraying the test as not yielding group differences, Spencer et al., 1999), (2) rendering it irrelevant (e.g., portraying the test as nonevaluative, Steele & Aronson, 1995), and (3) providing an identity-relevant antidote to stereotype threat (e.g., a value-affirmation, Martens, Johns, Greenberg, & Schimel, 2006). The results were clear and compelling. Under conditions that reduce threat, stereotyped students

performed better than nonstereotyped students who had the same level of prior real-world grades or test scores. The magnitude of this superior performance indexes the degree to which the prior measures underestimated stereotyped students' ability. It was just under one-fifth of a standard deviation ($d = 0.18$). Notably, this mean-level bias was found even though the prior performance measures predicted subsequent performance, and did so equally well for all groups. Psychological threat undermined stereotyped students' level of performance; it did not render that performance nonpredictive (see also Figure 1).[2]

A striking finding involved the ubiquity of the bias. It was found among students with all levels of prior performance, from diverse stereotyped groups (e.g., various ethnic minorities, girls and women), of diverse ages (K-12 through college) and nationalities (US and non-US), and in studies that used a wide range of measures of prior performance (various real-world tests and grades). None of these variables moderated the effect. In addition, the included studies featured diverse postmanipulation measures of performance (various intellectual tests) and diverse strategies to reduce threat. The broad range of populations and testing conditions in which latent ability was found suggests that psychological threat is the norm in academic environments (Steele, 1997).

The third and perhaps most meaningful test of the hypothesis that stereotype threat undermines students' real-world performance comes from randomized interventions to reduce stereotype-related threat in field settings. These interventions have been shown to raise stereotyped students' scores on standardized tests (e.g., Good, Aronson & Inzlicht, 2003) and grades in real-world classrooms (e.g., Cohen, Garcia, Apfel, & Master, 2006; Steele et al. 1998; Walton & Cohen, 2011) and to narrow group differences. In a later section we describe these interventions in detail. They provide rigorous causal evidence that stereotype-related threat

undermines students' real-world academic performance.

In addition, these interventions provide an opportunity to test whether the latent ability effect replicates when the performance outcome is students' real-world grades. In a second meta-analysis, Walton and Spencer (2009) examined three randomized interventions designed to reduce stereotype-related threat among Black students (n = 15,796; Cohen et al., 2006; Steele et al., 1998; Walton & Cohen, 2007). The results paralleled those from the laboratory studies. Black students in intervention conditions performed significantly better than White students who had the same prior grades and test scores. Once again this bias was found throughout the performance range (see Figure 1), and the effect size (d = 0.17) nearly matched that from the laboratory studies (d = 0.18). The probability of observing both effects by chance alone was 5 in 1 million. These results provide compelling evidence that the real-world performance of ethnic minorities in environments low in stereotype threat is underestimated by their test scores and grades as assessed in typical environments.

The size of the effect observed in the latent ability meta-analyses—$0.17 \leq d \leq 0.18$—almost certainly *under*estimates the degree of bias in common performance measures. The effect size reflects the amount by which threat was reduced in the postmeasure of performance relative to the premeasure of performance. If some threat persisted in conditions designed to reduce threat, the observed effect size underestimates the degree of bias on the prior measure.[3] In light of its ubiquity and magnitude, we believe that this bias, even when unintended, is one of the most insidious forms of bias in modern school and work environments.

The meta-analyses directly test the concern about stereotype threat raised by Sackett and colleagues (2004, 2008). These scholars assert that stereotype threat is only a laboratory phenomenon—that it undermines the performance of stereotyped students below the level that

would be expected on the basis of their prior performance. Consistent with this view, the meta-analysis of laboratory experiments yielded evidence of *underperformance*: At the same level of prior performance, stereotyped students performed worse in threatening conditions than nonstereotyped students. In the purified conditions of the laboratory, it may be that the level of threat can be increased above the level of threat in real-world settings (for alternative explanations, see Cohen & Sherman, 2005 and Footnote 3). But critically, the meta-analysis also yielded evidence of latent ability: In *non*threatening conditions, at the same level of prior performance stereotyped students performed better than nonstereotyped students. Sackett and colleagues may be right that a portion of the stereotype threat condition difference is due to the high level of threat that can be created in the laboratory. But this does not explain the whole condition difference. Rather, the stereotype threat condition difference also occurs because, in conditions that reduce threat, stereotyped students do *better* than expected based on their prior performance. That prior performance was itself polluted by stereotype threat. It underestimated stereotyped students' ability and potential in an environment without threat.

How Biased Are Real-World Measures of Academic Performance?

The meta-analytic results suggest that psychological threat biases standard measures of academic performance. How large is this bias? How much of real-world group differences does threat account for? We illustrate the size of the effect using the SAT, as this was the most common measure of prior performance in the studies included in the meta-analyses.

Extrapolating the results to the SAT raises the question of how representative the meta-analytic samples are of SAT testtakers. The samples are surely not fully representative but there are reasons to believe that the observed effect sizes generalize to broader samples and thus provide a reasonable first estimate of the degree of bias on the SAT. First, the effect was

unmoderated by type of premeasure (e.g., standardized test scores vs. classroom grades), by demographic variables (e.g., type of stereotyped group), and by level of prior performance (i.e., the slopes in the different cells were parallel; see Figure 1). Second, although some studies targeted populations that are more affected by negative stereotypes—such as students personally invested in the domain at hand, (Steele, 1997)—and could thus overestimate the effect, others recruited students from broad, heterogeneous populations. Third, as noted, the observed effect sizes may minimally estimate the bias. As researchers develop more effective ways to reduce threat, the estimate may rise. Consistent with this hypothesis, there was a positive relationship between the size of the latent ability effect observed in the laboratory experiments and the year in which the study was completed. More recent studies yielded larger effects ($r = .38$, $p = .016$). Among the most recent half of studies (those conducted in 2004 or later), the latent ability effect size was $d = 0.25$ ($Z = 3.90$, $p = .0001$).

With these considerations in mind, we estimate the degree of bias on the SAT, emphasizing that these are initial estimates and that future research may provide more precise estimates. Given the differences in populations and the possibility that the observed effect sizes underestimate the bias, the effect sizes (ds = 0.17 and 0.18) are best seen not as point estimates but as empirical guidelines to the likely degree of bias. We therefore provide a range of estimates: a low estimate slightly smaller than the degree of bias observed in the meta-analyses ($d = 0.15$) and two slightly higher estimates (ds = 0.20 and 0.25). Table 1 displays the proportion of group differences on the SAT that result from psychological threat given these estimates. The estimates suggest that psychological threat accounts for 57-94% of the gender gap on the SAT-Math test and 23-39% of the White/Latino gap and 17-28% of the White/Black gap on the SAT. Of course, many other factors contribute to racial/ethnic differences in academic performance,

especially poverty (Phillips, Brooks-Gunn, Duncan, Klebanov, & Crane, 1998). As the table shows, together socioeconomic status (SES) and psychological threat may account for much of race differences on the SAT.

Affirmative Meritocracy Step 1: Create a Stereotype-Safe Environment

In preventing people from performing to their potential, psychological threats undermine meritocracy. The first step to restore meritocracy is to create *stereotype-safe environments* that minimize such threats. Doing so would allow schools and employers to recover significant human potential.

An important question is whether organizations should first test whether psychological threats depress students' or employees' performance before working to create stereotype-safe environments. We suggest that they need not for two reasons. First, in many ways efforts to remove stereotype-related threat are simply good educational and organizational practice, which benefit all people even as they are especially beneficial to members of stereotyped groups.[4]

Second, although evidence suggests that psychological threats are pervasive in intellectual settings, it may be difficult to determine the level of threat in a specific context. For instance, while individual-difference measures can provide a measure of the level of threat experienced in a setting (Cohen & Garcia, 2005; Mendoza-Denton et al., 2002; Pinel, 1999), people may not have full or unbiased access to these internal processes (Schmader, Johns, & Forbes, 2008). Members of disadvantaged groups may be motivated to see the world as fair (Laurin, Fitzsimons, & Kay, 2010; Lerner, 1980), to deny prejudice they experience (Crocker & Major, 1989; Taylor, Wright, Moghaddam, & Lalonde, 1990), and to suppress thoughts of negative stereotypes (Logel, Iserman, Davies, Quinn, & Spencer, 2009). Further, any pattern of relative performance could occur when psychological threats are present. If an organization

observes underperformance—if members of a stereotyped group perform worse than others with the same level of prior performance—then psychological threats (but possibly other factors) likely depress the performance of members of the stereotyped group below their potential. The absence of underperformance suggests that the current environment is not more threatening than the prior environment but the two environments may be threatening to the same degree. Even when latent ability is observed, threat may be present if the prior environment was especially threatening. In all these cases, reducing threat further would raise the performance of people from stereotyped groups.

A demand for proof of the presence of psychological threat in a setting may only delay action. Consider the confession of S. J. Green, director of research for British American tobacco, who, after years of denying the evidence that smoking causes cancer, rejected obstructionism "not just morally but intellectually" (Oreskes & Conway, 2010, p. 274). Green wrote:

> A demand for scientific proof is always a formula for inaction and delay, and usually the first reaction of the guilty. The proper basis for such decisions is, of course, quite simply that which is reasonable in the circumstances.

When there is good reason to suspect that psychological threats undermine the performance of students or employees, and when an organization can take relatively simple steps to address this possibility, it should do so.

The case for action is especially strong where factors exist that research identifies as increasing the likelihood that people will experience stereotype threat. These include settings in which people work on difficult tasks that are represented as evaluative of abilities associated with social stereotypes (Steele & Aronson, 1995); cues that make group identity seem to be a potential basis of evaluation or exclusion, such as requests to report one's gender or ethnicity

(Steele & Aronson, 1995) or ambient cues that seem to define a field by group identity (Cheryan et al., 2009); the numeric underrepresentation of negatively stereotyped social groups (Inzlicht & Ben-Zeev, 2000; Murphy et al., 2007; Purdie-Vaughns et al., 2008), especially among people in power; sexist or racist treatment (Logel et al., 2009b); and messages about diversity that do not acknowledge or value the positive characteristics and contributions of people from diverse groups (e.g., color-blindness ideologies; Purdie-Vaughns et al., 2008; see also Derks, van Laar, & Ellemers, 2007.

How can a school or company create a stereotype-safe environment? Research suggests a number of strategies, which we review below. But there is no magic bullet. The nature of the threat may differ by context (see Shapiro & Neuberg, 2007). Organizations should test different strategies to identify those that are most effective for them. Additionally, we emphasize that psychological threat is only one cause of achievement gaps. Structural factors alone, such as high levels of poverty in ethnic minority communities, can cause large group differences (Phillips et al., 1998). In addition, students in low-income, minority classrooms may literally receive a gap-producing education in curricula and pedagogy (e.g., Darling-Hammond, 1990; Ladson-Billings, 2006). Interventions to reduce psychological threat do not mitigate poverty, improve teacher quality, or teach academic content. Instead, they allow students to take advantage of learning opportunities available to them (Yeager & Walton, 2011). In this way, threat-reduction, poverty-reduction, and curricular and pedagogical reforms are complimentary and nonsubstitutable. All are needed to close achievement gaps.

Create a Genuinely Welcoming Environment

The first step to reducing psychological threat is to ensure that the environment is genuinely welcoming of people from diverse groups. Obviously prejudice does not promote a

positive environment; if prejudice—whether conscious or nonconscious—exists in a setting, it should be reduced (Kang & Banaji, 2006). Indeed, teachers' level of implicit racial bias predicts the size of the racial achievement gap in their classrooms (Van den Bergh, Denessen, Hornstra, Voeten, & Holland, 2010). Even subtle cues communicated as a result of prejudice can have negative effects (Sue et al., 2007). In one series of studies, female engineering students interacted with male students who had previously completed a subtle measure of sexism. More sexist men behaved more dominantly and women they interacted with scored worse on a subsequent engineering exam (Logel et al., 2009b).

Incidental cues too can communicate to targets of stereotypes that they are not valued. Exposure to gender-stereotypic commercials (like for skin care products) can undermine women's aspirations and performance in math and science (Davies et al., 2002). Ambient cues that prime a masculine representation of a field, like Star Trek posters and video games in a computer science setting, can undermine women's sense of belonging and interest (Cheryan et al., 2009). A company that espouses colorblindness but employs few ethnic minorities may seem untrustworthy to Black professionals (Purdie-Vaughns et al., 2008; see also Murphy et al., 2007). Changing the representations people are exposed to (e.g., replacing Star Trek posters with nature posters), endorsing a multicultural philosophy that explicitly values diversity, conveying that the unique qualities of people from minority groups such as their language or religion are valued (Derks et al., 2007), or increasing the representation and visibility of minority-group members can thus improve stereotyped students' outcomes.

In testing settings too, incidental cues can cause stereotype threat if they lead students to suspect that their group identity is at risk (see Steele et al., 2002). Removing these cues can raise performance. One such cue is the request to report one's race or gender before a test. In a classic

study, Steele and Aronson (1995, Study 4) found that Black students asked to indicate their race before an ostensibly nonevaluative test performed worse. Presumably, the request evoked in Black students a worry about being viewed in light of the stereotype. Do such demographic queries add to the burden of stereotype threat in real-world settings where tests are represented as evaluating ability? In some cases they can. In a reanalysis of field-experimental data, Danaher and Crandall (2008) found that requesting demographic information after instead of before the AP calculus test raised girls' scores. If implemented nationwide, Danaher and Crandall estimated that this change would increase the number of girls who receive college calculus credit each year by 4,700. The standardized testing industry disputes this conclusion (Stricker & Ward, 2008), and there remain questions about when and for whom demographic queries exacerbate threat on evaluative real-world tests. But insofar as such a change is essentially cost-free and could bring important benefits, it is reasonable to make.

Another cue that can cause threat is working in the presence or numeric majority of members of the nonstereotyped group. Taking a math test in a group of men, for instance, can trigger stereotype threat among women and undermine performance (Inzlicht & Ben-Zeev, 2000). One field study with a national sample found that Black adults performed worse when a vocabulary quiz was administered in a face-to-face conversation by a White interviewer than by a Black interviewer (Huang, 2009). The Black interviewer cut the gap in performance between Black and White respondents by half and, controlling for demographic variables, eliminated it. Similarly, taking advantage of natural experiments, large-scale field studies find that having a female professor can raise women's performance in math and science classes (Carrell, Page, & West, 2010) and having a same-race teacher can raise Black students' school achievement (Dee, 2004; see also Massey & Fischer, 2005). This research suggests that, in some cases, altering the

social composition of academic settings might reduce threat. That said, it may be impractical in many settings to rearrange the social environment. There may also be political and ethical reasons to choose alternative strategies to reduce threat (e.g., fostering better intergroup relationships, Walton & Carr, 2012).

Create a Positive Subjective Environment

Creating a genuinely welcoming environment may be necessary but insufficient. A key insight in past research involves the importance of *subjective construal* (Steele, 1997). Two people can construe and experience the same event very differently. As only stereotyped students risk being viewed in light of the stereotype, taking an evaluative test may be threatening to stereotyped students but not to others. Subjective construal provides an important point of entry for intervention. Changing stereotyped students' construal of intellectual tests and other aspects of the academic environment can boost their motivation and performance.

Directly challenge students' assumptions about intellectual tests. The perception that an intellectual test is evaluative of a stereotyped ability can trigger stereotype threat (Steele & Aronson, 1995). In real-world settings, it may not be possible (or ethical) to portray high-stakes tests as nonevaluative. But threat can also be mitigated by refuting the validity of the stereotype. For instance, assuring people that a test is "fair" across different groups —that it yields no group difference in performance—can make tests, in actuality, fairer (Spencer et al., 1999). In one study, White women scored 50% better on a practice calculus final exam said to be "fair" across gender groups as compared to when it was presented merely as evaluative of ability (Good, Aronson & Harder, 2008). We do not advocate misrepresenting tests. But whether or not efforts are made to change students' representations, students take tests with assumptions about what the test evaluates and how they and their group stack up relative to others. If group differences

emerge primarily *because* people anticipate such differences, it is important to challenge this assumption.

Indirect strategies to reduce apprehension about negative stereotypes. The worry that one could be perceived in light of a negative stereotype can also be allayed indirectly. For instance, asking students to describe aspects of their individual self before a test can convey that they are seen as individuals rather than as representatives of a group, and thus improve performance (Ambady, Paik, Steele, Owen-Smith, & Mitchell, 2004; Gresky, Ten Eyck, Lord, & McIntyre, 2005). Likewise, blurring the perceived boundaries between groups, as by asking women to reflect on characteristics that are shared between women and men, can make group stereotypes seem less relevant (Rosenthal & Crisp, 2006; see also McGlone & Aronson, 2007).

Another way to reduce apprehension about negative stereotypes is to facilitate better, relationships between people from stereotyped and nonstereotyped groups (Walton & Carr, 2012). Insofar as stereotype threat arises from a worry about how one will be perceived and evaluated across group lines (e.g., Cohen & Steele, 2002; Logel et al., 2009b), improving intergroup relationships may reduce worries about being evaluated negatively. For instance, one series of studies found that small cues that created the sense of working with a man on a math test reduced stereotype threat among women even when students performed individually (Carr, Walton, & Dweck, 2012). In this research, receiving a friendly tip from a man taking the same test raised women's scores on an evaluative math exam; receiving the same tip from a "tip bank" had no effect. The effect was mediated by women's perception of how the man regarded them. Similarly, in field settings, cooperative learning programs like the "jigsaw classroom" simultaneously improve intergroup relationships and raise minority students' achievement (Aronson & Osherow, 1980). Additionally, structured activities to facilitate cross-race

friendships can improve minority students' experience in the transition to predominantly White universities (Mendoza-Denton & Page-Gould, 2008).

These strategies reduce the perceived relevance of negative stereotypes—they make it seem less likely that one will be viewed through the lens of a negative stereotype. Indirect strategies can also invalidate stereotypes. For instance, exposure to in-group role models in a threatened domain, such as to a woman who is skilled at math, can discredit the stereotype and reduce threat on evaluative tests (Marx & Goff, 2005; Marx & Roman, 2002; McIntyre, Paulson, & Lord, 2003).

Help students cope with stereotype threat: Reattribution, reappraisal, and retraining. Another way to mitigate stereotype threat is to give students tools to cope with threat they experience. For instance, teaching stereotyped students about stereotype threat can lead students to attribute anxiety or physiological arousal they experience while taking an evaluative test to stereotype threat rather than to a risk of failure and thus improve performance (Johns, Schmader, & Martens, 2005). More broadly, teaching students to reappraise their emotional reactions to a test—for instance, to see the test in an objective manner rather than as personally relevant, or to view anxiety as a potential source of strong performance rather than as a hindrance—can improve stereotyped students' scores (Johns, Inzlicht, & Schmader, 2008). Relatedly, one source of the performance decrements associated with stereotype threat involves people's efforts to suppress negative thoughts and emotions about stereotypes as they take a test, which consumes limited cognitive resources. Encouraging less costly coping strategies can raise performance. One study found that asking women to replace feelings of worry with thoughts of a neutral object (e.g., "a red Volkswagen") raised women's scores on an evaluative math exam and eliminated gender differences (Logel et al., 2009a). Finally, retraining people's associations to disconfirm

negative stereotypes can raise scores (Forbes & Schmader, 2010). In one study, an associative training task that led women to pair the concepts "women are good at" and "math" improved women's math performance in the face of threat a day later. However, the robustness of this approach in settings that reinforce stereotypes is not known.

So far this section has emphasized laboratory research and interventions that target specific performance opportunities. But brief interventions—some lasting an hour or less—can also reduce psychological threat in school settings broadly, and raise stereotyped students' academic performance over months and even years. How is this possible? Social-psychological interventions replace negative self-perpetuating processes that undermine students' outcomes over time with positive processes (Yeager & Walton, 2011). For instance, a secure sense of belonging or an adaptive construal of critical feedback may help students form the kinds of relationships in school needed to support intellectual growth and high performance over time.

Buttress students' sense of social belonging. Students who face negative stereotypes in school may reasonably wonder if others will fully include and value them (Walton & Cohen, 2007). As a result, stereotyped students may view negative social events in school (e.g., social exclusion) as evidence that they do not belong in school in general. This inference saps motivation. To prevent such deleterious attributions, *the social-belonging intervention* provides students a nonthreatening explanation for negative social events in school—it leads students to see these events as a normal and temporary part of people's experience in school. In one study, first-year college students read a survey of upper-year students at their school (Walton & Cohen, 2007, 2011). The survey indicated that negative social events and feelings of nonbelonging are normal at first in college (e.g., experienced by students of all ethnicities) and dissipate with time. The treatment was designed to lead students to attribute such events to the difficulty of the

transition to college rather than to a lack of belonging. The treatment message was reinforced using "saying-is-believing" exercises—for instance, students wrote an essay about how they had experienced this process of change, ostensibly to be shared with future students to improve their college transition—a powerful persuasive tactic. In total, the intervention lasted about 1 hour.

For White students, who have little cause to doubt their belonging in school on account of their race, the treatment had little effect. But in two cohorts of students and relative to multiple control groups, the intervention raised Black students' GPA over the subsequent three years of college, reducing the Black-White gap in achievement over this period by 50% (Walton & Cohen, 2011). Daily diary measures suggest that the intervention worked through the intended psychological process. In the week immediately following its delivery, the intervention prevented Black students from perceiving daily instances of adversity as evidence of a global lack of belonging; this mediated the long-term gain in GPA. In addition, at the end of college Black students in the treatment condition reported feeling more secure in their belonging on campus, evidenced less thought about negative racial stereotypes, and exhibited other benefits of a secure sense of belonging, including better health and greater happiness. Related interventions have been found to raise women's achievement in engineering (Walton, Logel, Peach, Spencer, & Zanna, 2012) and Black students' achievement in middle school (Walton, Cohen, Garcia, Apfel, & Master, 2012).

Encourage adaptive construals of critical feedback. One context that can trigger threat but is especially important for learning and growth is the receipt of critical feedback. For stereotyped students, critical feedback can be ambiguous in meaning, especially when received across group lines. It could reflect an honest assessment of one's performance and provide valuable information about ways to improve. Or it could result from bias. Laboratory research

finds that disambiguating critical feedback—telling students that they are receiving critical feedback because of the high standards of the task and because the feedback-giver believes in their potential to meet those standards—reduces perceptions of bias among stereotyped students and sustains their task motivation (Cohen & Steele, 2002; Cohen, Steele, & Ross, 1999). Based on this research, an intervention used "saying-is-believing" exercises to lead high school students to interpret critical feedback in general as a sign of teachers' care and high standards. This intervention raised Black students' GPA the next semester and reduced the racial achievement gap by 40% (Yeager et al., 2011). These findings also suggest a second intervention strategy: Perhaps training teachers, mentors, and supervisors to disambiguate critical feedback as they provide it would raise stereotyped students and employees' motivation and performance.

Buttress students' sense of self-efficacy. A related intervention aims to prevent students from viewing academic setbacks as global evidence that they cannot succeed. For instance, students may be taught that academic struggles are normal in the transition to a new school and lessen with time (Wilson, Damiani, & Shelton, 2002) or that intelligence is malleable and that, with effort, they can overcome setbacks and master challenges (Blackwell, Trzesniewski, & Dweck, 2007). These approaches can raise academic performance in the face of difficulty for all students but may be especially effective for students who confront stereotypes that allege fixed inability (Aronson et al., 2002). In one study, 7th grade girls exposed to mentors who delivered one of several such interventions scored better on a statewide standardized math test at the end of the school year than girls in a control condition (Good et al., 2003). Boys showed marginal benefits, and the interventions eliminated gender differences in math scores.

Value-affirmations to reduce stress and threat. A final intervention aims to buttress students against threat. Decades of laboratory research show that providing people opportunities

to reflect on personally important values can bolster their sense of self-integrity—a view of themselves as good, virtuous, and efficacious—and, as a result, reduce psychological stress and threat (Sherman & Cohen, 2006). Threats to a specific identity or aspect of the self simply feel less threatening when one's overall sense of self-integrity feels secure (Sherman & Hartson, 2011). These "value-affirmations" can improve stereotyped students' performance in laboratory (Martens et al., 2006) and real-world settings. In one field-experiment, White and Black 7th graders completed a value-affirmation as an in-class writing exercise (Cohen et al., 2006). Students identified their most important values from a brief list and wrote about the importance of these values to them. The exercise aimed to remind students of unconditional sources of worth in a potentially threatening environment. Control students identified their least important values and wrote about why they might matter to someone else. The affirmation had no effect on White students but it reduced the accessibility of racial stereotypes among Black students and raised their end-of-term course grades by one-third of a grade-point, reducing the racial achievement gap by 40%. Long-term follow-ups with three cohorts of students showed that the boost in GPA for Black students persisted over the last two years of middle school, apparently by interrupting a negative recursive cycle whereby poor performance begat worse performance over time (Cohen, Garcia, Purdie-Vaughns, Apfel, & Brzustowski, 2009). Subsequent trials have shown value-affirmation interventions to raise achievement among Latino middle school students (Sherman et al., under review) and women in college physics courses (Miyake et al., 2010).

Implementation in an organizational context: The 21st Century program. How can organizations implement threat-reducing interventions? The 21st Century program at the University of Michigan provides one model (Steele, 1997; Steele et al., 1998). The program—an ethnically diverse dormitory for first-year students—combined several key elements. First, the

program was presented as honorific not remedial. Although remedial programs may sometimes have value, they risk triggering in stereotyped students the very stereotype that impugns their ability. Instead, the program communicated the university's high regard of and high standards for the participating students (Cohen & Steele, 2002). Second, the program featured weekly discussion groups in which students learned that difficulties they experienced adjusting to college were shared and the product of challenges faced by all students rather than indicative of a lack of belonging (Walton & Cohen, 2011). The program had little effect on White students but it reduced self-reported levels of stereotype threat among Black students, increased Black students' identification with school, and raised their first-term GPA by one-third of a grade point.

Summary

The research reviewed here shows that even brief interventions to mitigate stereotype-related threat can generate large, long-lasting benefits for ethnic minority students and women, improving test scores and grades months and years later. These interventions teach no academic content or abilities. Instead, they create stereotype-safe environments—they reduce psychological threats present in school and allow stereotyped students to exhibit intellectual capabilities that are already present in them. In removing psychological threats, these interventions complement traditional educational reforms that ensure that all students have access to high-quality educational experiences.

The effectiveness of the interventions described here highlights the presence of stereotype-related threat in common academic environments and the benefits of removing these threats. However, questions remain about what form the threat takes in different settings and for different populations and when and with whom specific interventions will be most effective. These are critical questions for future research. This research should take the form of local

experimentation in which researchers and organizations work together using a field-experimental methodology to identify the specific form of threats present in particular school and work environments and to develop and evaluate interventions to mitigate these threats (Yeager & Walton, 2011). Especially important is research in contexts that have characteristics that heighten the likelihood that people will experience stereotype threat. This research may shed light on important theoretical and applied questions, including the mechanisms through which social-psychological interventions work, how their effects persist over time, when and for whom they are most effective, and how they can be scaled-up to larger and more heterogeneous populations. This work is essential if threat-reducing interventions are to fulfill their promise to significantly narrow achievement gaps on a large scale.

Finally, we note that federal law permits organizations to implement, evaluate, and improve on threat-reducing interventions. When state entities draw classifications or confer benefits on racial lines, courts subject their policies to significant scrutiny (e.g., *Gratz v. Bollinger*, 2003). The proposed interventions do neither, in part because making group identity salient can trigger stereotype threat (Steele & Aronson, 1995). They thus benefit students of all groups—not just women and racial minorities but anyone for whom relevant psychological barriers impede performance (e.g., Aronson et al., 2002). In these respects, the interventions are constitutionally favored ways to remedy bias and promote excellence (*Gratz v. Bollinger*, 2003; *Grutter v. Bollinger*, 2003; Primus, 2003). This said, the interventions do not ignore race. To mitigate psychological threats it is necessary to consider their impact on relevant groups, including racial groups. One cannot be blind to race in developing interventions to reduce threat but these interventions are nonetheless race neutral in aim and operation. Lastly, if organizations were to demonstrate the effectiveness of affordable interventions that mitigate threats that

disproportionately harm women or minorities, federal agencies could feel obligated to use their powers under Title VI of the Civil Rights Act of 1964 to require recipients of federal funds (most schools; many state agencies, businesses, and nonprofits) to adopt them (Abernathy, 2006).

Affirmative Meritocracy Step 2: Account for Latent Ability in Admissions and Hiring

Suppose a school or company creates a stereotype-safe environment and observes evidence of latent ability. Perhaps female science students earn better grades than male science students despite having earned the same score on an entrance exam. Perhaps ethnic minorities do better on the job than their nonminority peers—billing more hours at a law firm, or selling more real estate. The observation of latent ability suggests that the prior measure underestimates the ability and potential of stereotyped individuals. If so, how should the organization interpret that measure in making admissions or hiring decisions? The question is pressing when the measure, although showing a mean-level bias, predicts school or work performance for both groups. It thus provides valuable information about which prospective students or employees are likely to be most successful and productive even as it underestimates the likely performance of members of one group relative to members of another group.

The Ideal Remedy

The ideal remedy is to reduce psychological threat on the prior measure so that scores on it predict the same level of subsequent performance for all groups. A fundamental lesson of stereotype threat is that situational factors impair performance; situational remedies to reduce threat are thus ideal . But an organization may not control the environment in which the prior measure was assessed. Colleges do not administer the SAT; companies do not run colleges. And performance measures may derive from many contexts—such as GPAs earned by students in diverse high schools—so there is not one environment to fix but many. In the long-term,

reducing threat will require the collective efforts of many schools, employers, testing organizations, and researchers.

Alternative Remedies

If reducing threat on the prior measure is outside an organization's control, an exceedingly difficult question of social policy, psychology, and law arises. It would be inappropriate to accept a measure assessed in a biased setting at face value to make selection decisions. To do so would be to discriminate against stereotyped groups—to reject more qualified people from stereotyped groups in favor of less qualified people from nonstereotyped groups. Selection procedures should be remedied to remove this bias. But how? Below we outline five potential remedies, one with two variants. Each offers different advantages and drawbacks, and reasonable people may disagree about their relative merits in different circumstances. Our purpose is not to advocate for any one remedy but, rather, to suggest several alternatives. We emphasize that, when bias is observed, organizations must consider these alternatives (and potentially others) to make meritocratic, nondiscriminatory selection decisions.

1. Do not use measures assessed in biased settings. One approach is to forego measures assessed in biased settings or, at least, to reduce their weight in decision-making. Because these measures are, to an organization, simply biased, as shorthand we refer to them as biased measures. For instance, the Society for Industrial and Organizational Psychology (2003) writes, "In general, the finding of concern would be evidence of substantial underprediction of performance in the subgroup of interest. Such a finding would generally preclude operational use of the predictor [in selection decisions]" (p. 46). Evidence of bias is certainly good cause not to use a measure. If a law school found that LSAT scores are biased predictors of law-school performance, it might seek alternative measures to make admissions decisions. We address

alternative measures below. This remedy may be most useful, however, when a particular performance measure is biased, for instance as a result of bias in content. When indicators of merit are biased generally—for instance, as a result of bias in common performance contexts—this remedy may seem to preclude the use of *any* performance measure in selection decisions. In addition, simply foregoing biased measures is a blunt solution: If the measure, although biased against stereotyped groups, predicts subsequent performance, not using it may entail ignoring potentially valuable information. As a consequence, this approach may be less appealing when the biased measure is highly predictive of outcome measures (and less biased measures are unavailable), and more appealing when the measure is less predictive (and less biased measures are available).

2. Replace measures assessed in biased settings. A measure assessed in a biased setting may be replaced with a measure assessed in a less biased setting. For instance, standardized tests administered in contexts designed to mitigate psychological threat could provide a preferable alternative to tests administered in traditional contexts, even when the tests have the same content. This approach complements efforts to widen the range of abilities assessed. As noted, Sternberg and colleagues (2006) showed that assessing a broader range of college applicants' skills than is typical can simultaneously shrink group differences and raise correlations with college grades. An ideal measure would both assess an appropriately broad range of abilities and be assessed in a low-threat setting.

3. Restrict the role of measures assessed in biased settings. In several ways measures assessed in biased settings can be retained but restricted in their role in decision-making. These approaches, however, are problematic. For instance, an admissions committee could ignore the continuous nature of applicants' test scores and instead determine only whether an applicant's

score surpasses a given cut off or not (e.g., a 1200 SAT-Score, top 20% of applicants). While this approach might reduce the contribution of a biased measure to selection decisions, it does not solve the problem at hand: In undermining stereotyped students' mean scores, psychological threat will also reduce the number whose scores surpass any given cut off.

Alternately, an admissions committee might evaluate candidates from different social groups separately. For instance, a school could admit candidates whose test scores fall in the top 20% of applicants from their racial-ethnic or gender group. This approach raises a number of problems. By encouraging only within-category comparisons, it risks reifying social categories. It also assumes that there are not real differences in developed intellectual ability and potential between social groups, which may not be the case. And it raises seemingly insurmountable legal concerns, for the Supreme Court has repeatedly struck down selection schemes that use racially separate tracks to insulate applicants from one racial group from competition with applicants from other racial groups (e.g., *Grutter v. Bollinger*, 2003).

4a. Correct scores to reduce bias and promote merit. A fourth approach is to retain the full range of the biased measure but to compensate stereotyped individuals for the bias observed in it. In this approach, after reducing threat in their internal environment, organizations would assess the size of the bias in a given selection measure (i.e., the size of the latent ability effect) observed within their local setting. They would then add to stereotyped people's scores the number of points that correspond to the magnitude of the observed bias. If an organization observed a bias of one-fifth of a standard deviation, it would increase the scores of stereotyped people on the relevant measure by that amount. We describe this approach in detail as it raises complex questions.

Score corrections offer the advantage of remedying a group-level bias at the level of the

group while preserving the full predictive value of the performance measure. They ensure that scores reflect the same level of ability and potential for people from all groups. As such, they may be more attractive when measures are more predictive of outcomes. Score corrections have drawn support from diverse scholars. Sackett and Wilk (1994) call "score adjustment[s]" "a technically appropriate solution" (p. 933) for a finding of latent ability. In *The Bell Curve*, Herrnstein and Murray (1994) write (p. 280):

> If the SAT is biased against Blacks, it will underpredict their college performance … It would be as if the test underestimated the 'true' SAT score of the Blacks, so the natural remedy for this kind of bias would be to compensate the Black applicants by, for example, adding the appropriate number of points onto their scores.

Score corrections differ from other score modifications in that they are specifically justified by merit—they are score *corrections*. To ensure that they promote merit, score corrections are grounded in local empiricism—by the finding of latent ability of a specific size in a specific group in a specific organization on a specific performance measure. In this context, correcting scores would increase the diversity of selection decisions *by* promoting merit and would also thereby raise organizational performance. Odds are that candidates who benefit would simply be better candidates—more able, with greater potential—than candidates they displace. In addition, score corrections could respond dynamically to changing circumstances: As an organization reduces threat more, it may observe evidence of greater bias (e.g., a larger latent ability effect) and implement larger score corrections; as threat in the prior performance environment is reduced, the organization may observe less bias and use smaller score corrections. In this way, score corrections are self-extinguishing. Widespread reduction of threat

in society will reduce bias in performance measures and the need for score corrections.

Score corrections mitigate some concerns about traditional forms of affirmative action. One such concern is whether affirmative action harms stereotyped students by placing them in settings where they struggle to compete (e.g., Sander, 2004b). With score corrections, this would not be the case. By reducing threat in the internal environment first and introducing score corrections only as a function of the degree of bias observed, stereotyped students should be as likely as other students to perform well. A second concern is whether affirmative action reinforces negative stereotypes and/or undermines confidence in the self (Crosby et al., 2003; Crosby 2004). Properly communicated, score corrections should allay these problems. Score corrections are not "free points" but empirically validated corrections for biased measures and antidotes to discrimination.[5]

Despite these advantages, important concerns may constrain the appeal and value of score corrections. First, people may view score corrections negatively, for instance as "social engineering." Of course, all selection systems are socially engineered—they are not natural but constructed by people to advance specific aims (e.g., Crosby et al., 2003). In any selection system, people must decide what measures to use and how to interpret them. Still, important administrative questions such as who decides when to implement score corrections, how much bias is enough to introduce score corrections, who are candidates for score corrections, and when and how to modify score corrections as circumstances change may raise legitimate concerns.

Second, an important question involves to whom score corrections should extend. This review has focused on women and ethnic minorities as these are the groups most studied in research on stereotype threat. But what if other groups (e.g., religious minorities) or certain kinds of people (e.g., those with low self-esteem) did better in school or at work than their prior

performance would predict? One answer is that the principle of score corrections is a general one. If a measure substantially underpredicts later performance for any group or set of people the measure should be corrected. To fail to correct a biased measure would sanction discrimination, no matter the group. But the myriad ways in which people can be compared—along lines of race, gender, religion, age, social class, individual differences, etc.—limit the practical value of a purely empirical approach. A second answer is that biases based on social identity (e.g., stereotype threat), as opposed to personal identity (e.g., low self-esteem), may deserve special remedial action because of their ubiquity, their potential to harm not just individuals directly affected but also other ingroup members aware of that harm (Cohen & Garcia, 2005; H. R. Rep. No. 111-288, 2009), and because of the existence of a well-developed, empirically-substantiated theory that shows how, when, for whom, and by how much group-based threats undermine intellectual performance (Steele et al., 2002; Walton & Spencer, 2009).

4b. Correct scores on an individual basis. A third problem with group-based score corrections involves individual variability: Some students may experience more stereotype-related threat than others, for instance, as a result of different perceptions of threat (e.g., Mendoza-Denton et al., 2002; Pinel, 1999), different susceptibility to threat (e.g., women who are more invested in math are more vulnerable to stereotype threat; Nguyen & Ryan, 2008; Spencer et al., 1999), or different exposure to threat, for instance if they take tests in different settings or attend different classrooms. If so, score corrections may help some students "too little" and other students "too much." Score corrections correct for bias at the level of the group but not necessarily at the level of the individual. Of course, no assessment is perfect. Given current evidence, score corrections would improve the assessment of ability: They undo systematic bias and forestall group-based discrimination. But if accurate measures indexed

individual experiences of threat, organizations could create individuated score corrections that would remedy bias at both the level of the group and the level of the individual. Doing so would be far preferable. An important, open question is whether such valid and reliable measures could be developed, for instance based on existing individual-difference measures (Cohen & Garcia, 2005; Mendoza-Denton et al., 2002; Pinel, 1999), and if their validity and reliability could be maintained as they affect important selection decisions (e.g., Could people game the system by falsely reporting high levels of threat?). Importantly, the goal is not to perfectly assess the level of a threat a person has experienced—even in theory, any measure provides only a point estimate with associated error variance—but, instead, to reasonably approximate individual differences in the degree to which threat has undermined people's performance.

A final, important issue is that score corrections raise complex legal questions. On employment-related tests, mechanical race- and gender-based score adjustments and corrections are prohibited by Title VII of the Civil Rights Act of 1964 (as amended). In education and other contexts, courts applying Title VI of the Civil Rights Act of 1964 and the Constitution closely scrutinize actions taken by public entities and recipients of federal funds that differentiate people on the basis of race or sex. Here we focus on racial classifications. Sex classifications face an equally or more forgiving standard and so are as or more likely to pass legal muster. Mechanical race-based score corrections to promote merit pose special doctrinal problems. The Supreme Court has strongly disfavored mechanical race-based score adjustments based upon suppositions that, inter alia, they (1) are anti-meritocratic, (2) stigmatize, (3) promote racial hostility, and (4) perpetuate the salience of race (*Gratz v. Bollinger*, 2003; *Grutter v. Bollinger*, 2003; *Parents Involved in Community Schools v. Seattle School Dist. No. 1*, 2007; Primus, 2003). As compared to traditional mechanical race-based score adjustments, score corrections have very different

effects. They promote merit. As such, they should be less likely to stigmatize minorities or incite racial hostility. Indeed, people willingly endorse racial classifications that promote merit (Son Hing, Bobocel, & Zanna, 2002). To the extent that merit-based score corrections generate popular resentment, the Court must decide how much credence to give complaints by nonstereotyped individuals that a meritocratic policy deprives them of benefits they do not merit. The history of the civil rights movement in the United States is replete with instances in which federal courts and federal authority played crucial roles in countering and eventually reducing biases that had over time become accepted aspects of society. And score corrections, by allowing organizations to reap the benefits of reducing psychological threat, both encourage such reductions in threat and hasten the day when score corrections are unnecessary, thereby reducing the salience of race. Score corrections thus present the Court a potential choice between its uniform opposition to mechanical changes to scores on racial lines and many of the reasons it has articulated to justify that policy. Advances in science could resolve this question. If researchers developed a valid, reliable individual-difference measure of susceptibility to psychological threat that predicted latent ability, individuated score corrections would be legally and scientifically preferable.

5. Take bias into account in individualized selection processes. A fifth approach is to educate selection officers of the bias in specific performance measures and allow them to weigh this information in making individualized evaluations of candidates. Doing so would capture latent ability through individualized selection. The organization—a school or employer—would inform its selection officers that, on average, a measure underpredicts performance by members of a particular group by a particular amount. Taking this bias into account, those officers would then select candidates on the merits of their entire applications.

This approach raises legal issues. In brief, the Supreme Court permits institutions of higher education to take race or sex into account to promote diversity in the context of individualized evaluations of candidates for admission (*Gratz v. Bollinger*, 2003; *Grutter v. Bollinger*, 2003). The Court has not decided whether meritocratic decision making—the aim of the proposed approach—may also justify differentiating people by race or sex. Its opinions in related cases, however, emphasize the importance of merit (*Gratz v. Bollinger*, 2003; *Grutter v. Bollinger*, 2003; *Parents Involved in Community Schools v. Seattle School Dist. No. 1*, 2007 (plurality opinion); *Rice v. Cayetano*, 2000).[6] If merit does justify limited racial or sex classifications, then this policy, because it resembles the individualized selection process that the Court approved in *Grutter*, is the approach more likely be upheld. Further, if courts permit individualized selection processes to promote merit, federal agencies could feel obligated to use their powers under Title VI of the Civil Rights Act of 1964 to require recipients of federal funds to implement such measures (Abernathy, 2006).

Summary

The question of how to interpret a biased but predictive performance measure is exceedingly difficult and, we reiterate, people may disagree about the appropriate remedy. But we emphasize that, in the face of compelling evidence that measures of merit underestimate the ability and potential of members of stereotyped groups, the choice to do nothing is highly problematic: It would sanction discrimination against people from disadvantaged groups.

Conclusion

In 1807, in the midst of the Napoleonic Wars, the German mathematician Carl Friedrich Gauss received a letter from a long-time correspondent, the French mathematician, M. LeBlanc. To Gauss' shock, LeBlanc revealed that he was a she: Sophie Germain. Gauss' reply, in a letter

of April 30, 1807, is worth reflecting upon here:

> A taste for the abstract sciences in general and above all the mysteries of numbers is excessively rare … But when a person of the sex which, according to our customs and prejudices, must encounter infinitely more difficulties than men to familiarize herself with these thorny researches, succeeds nevertheless in surmounting these obstacles and penetrating the most obscure parts of them, then without doubt she must have the noblest courage, quite extraordinary talents and superior genius. (quoted in Bell, 2000, p. 333)

Germain was a brilliant mathematician but she hid her gender identity for years because she thought that a woman would not be taken seriously. What could Germain and her contemporaries have accomplished in a society that valued women in math? What could women and ethnic minorities accomplish today in school and work settings that value them?

A fundamental ideal in the United States and elsewhere is that all people, regardless of social background, have an equal opportunity to succeed. To further approach this ideal requires removing both structural obstacles to achievement and psychological barriers. Complementing traditional reforms, organizations should reduce psychological threat in their internal environments to ensure that all people can learn and perform to their potential. Further, in making admissions and hiring decisions, organizations should interpret measures of merit in ways that accurately index the ability and performance potential of all candidates. In taking these affirmative steps, organizations can promote meritocracy and diversity at once.

References

Abernathy, C. F. (2006). Legal realism and the failure of the "effects" test for discrimination. *Georgetown Law Journal, 94*, 267-319.

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.

Aronson, J. & Dee, T. (2012). Stereotype threat in the real world. In M. Inzlicht & T. Schmader (Eds.) *Stereotype threat: Theory, processes, and application* (pp. 264-279). Oxford, England: Oxford University Press.

Aronson, E. & Osherow, N. (1980). Cooperation, prosocial behavior, and academic performance: Experiments in the desegregated classroom. *Applied Social Psychology Annual, 1,* 163-196.

Aronson, J., Fried, C. B., & Good, C. (2002). Reducing the effect of stereotype threat on African American college students by shaping theories of intelligence. *Journal of Experimental Social Psychology, 38,* 113-125.

Bell, E. T. (2000). The prince of mathematics. In J. Newman (Ed.). *The world of mathematics: Volume 1* (pp. 295-339). Mineola, NY: Dover Publications, Inc.

Blackwell, L. S., Trzesniewski, K. H., & Dweck, C. S. (2007). Implicit theories of intelligence predict achievement across an adolescent transition: A longitudinal study and an intervention. *Child Development, 78,* 246-263.

Bobocel, D. R., Son Hing, L. S., Davey, L. M., Stanley, D. J., & Zanna, M. P. (1998). Justice-based opposition to social policies: Is it genuine? *Journal of Personality and Social Psychology, 75,* 653-669.

Bowen, W. G. & Bok, D. (1998). *The shape of the river*. Princeton: Princeton University Press.

Carr, P. B., Walton, G. M., & Dweck, C. S. (2012). *The feeling of collaboration forestalls stereotype threat*. Manuscript in preparation, Stanford University, Stanford, CA.

Camara, W. J. & Schmidt, A. E. (1999). *Group differences in standardized testing and social stratification* (College Board Research Report No. 99–5). New York: College Board.

Carrell, S. E., Page, M. P., & West, J. E. (2010). Sex and science: How professor gender perpetuates the gender gap. *The Quarterly Journal of Economics, 125,* 1101-1144.

Cheryan, S., Plaut, V.C., Davies, P., & Steele, C. M. (2009). Ambient belonging: How stereotypical environments impact gender participation in computer science. *Journal of Personality and Social Psychology, 97,* 1045-1060.

Cohen, G. L. & Garcia. J (2005). "I am us": Negative stereotypes as collective threats. *Journal of Personality and Social Psychology, 89*, 566-582.

Cohen, G. L., Garcia, J., Apfel, N., & Master, A. (2006). Reducing the racial achievement gap: A social-psychological intervention. *Science, 313,* 1307-1310.

Cohen, G. L., Garcia, J., Purdie-Vaughns, V., Apfel, N., & Brzustoski, P. (2009). Recursive processes in self-affirmation: Intervening to close the minority achievement gap. *Science, 324,* 400-403.

Cohen, G. L. & Sherman, D. K. (2005). Stereotype threat and the social and scientific contexts of the race achievement gap [response]. *American Psychologist, 60*, 270-271.

Cohen, G. L. & Steele, C. M. (2002). A barrier of mistrust: How stereotypes affect cross-race mentoring. In J. Aronson (Ed.), *Improving academic achievement: Impact of psychological factors on education* (pp. 305–331). Oxford, England: Academic Press.

Cohen, G. L., Steele, C. M., & Ross, L. D. (1999). The mentor's dilemma: Providing critical

feedback across the racial divide. *Personality and Social Psychology Bulletin, 25,* 1302–1318.

The College Board. (2011). *2011 college-bound seniors: Total group profile report*. Retrieved October 24, 2011, from http://professionals.collegeboard.com/profdownload/cbs2011_total_group_report.pdf.

Crocker, J. & Major, B. (1989). Social stigma and self-esteem: The self-protective properties of stigma. *Psychological Review*, *96*, 608-630.

Crosby, F. J. (2004). *Affirmative action is dead: Long live affirmative action*. New Haven, CT: Yale University Press.

Crosby, F. J., Iyer, A., Clayton, S., & Downing, R. A. (2003). Affirmative action: Psychological data and the policy debates. *American Psychologist, 58,* 93-115.

Cullen, M. J., Hardison, C. M., Sackett, P. R. (2004). Using SAT-grade and ability-job performance relationships to test predictions derived from stereotype threat theory. *Journal of Applied Psychology, 89,* 220-230.

Danaher, K. & Crandall, C. S. (2008). Stereotype threat in applied settings re-examined. *Journal of Applied Social Psychology, 38,* 1639-1655.

Darling-Hammond, L. (1990). Teacher quality and equality. In J. I. Goodlad & P. Keating (Eds.). *Access to knowledge: An agenda for our nation's schools*, pp. 237-258. New York: College Entrance Examination Board.

Davies, P. G., Spencer, S. J., Quinn, D. M., & Gerhardstein, R. (2002). Consuming images: How television commercials that elicit stereotype threat can restrain women academically and professionally. *Personality and Social Psychology Bulletin, 28,* 1615-1628.

Dee, T. S. (2004). Teachers, race, and student achievement in a randomized experiment. *The*

Review of Economics and Statistics, 86, 195-210.

Derks, B., van Laar, C. & Ellemers, N. (2007). The beneficial effects of social identity protection on the performance motivation of members of devalued groups. *Social Issues and Policy Review* 1, 217–256.

Forbes, C. E. & Schmader, T. (2010). Retraining attitudes and stereotypes to affect motivation and cognitive capacity under stereotype threat. *Journal of Personality and Social Psychology, 99,* 740-754.

Gardner, H. (1999). *Intelligence reframed: Multiple intelligences for the 21st century*. New York: Basicbooks.

Gonzales, P. M., Blanton, H., & Williams, K. J. (2002). The effects of stereotype threat and double-minority status on the test performance of Latino women. *Personality and Social Psychology Bulletin, 28,* 659-670.

Good, C., Aronson, J., & Harder, J. A. (2008). Problems in the pipeline: Stereotype threat and women's achievement in high-level math courses. *Journal of Applied Developmental Psychology, 29,* 17-28.

Good, C., Aronson, J., & Inzlicht, M. (2003). Improving adolescents' standardized test performance: An intervention to reduce the effects of stereotype threat. *Journal of Applied Developmental Psychology, 24,* 645-622.

Guinier, L. & Sturm, S., (2001). *Who's qualified?: A new democracy forum on creating equal opportunity in school and jobs*. Boston, MA: Beacon Press.

Gratz v. Bollinger, 539 U.S. 244 (2003).

Gresky, D. M., Ten Eyck, L. L., Lord, C. G., & McIntyre, R. B. (2005). Effects of salient multiple identities on women's performance under mathematics stereotype threat. *Sex*

Roles, 53, 703-716.

Grutter v. Bollinger, 539 U.S. 306 (2003).

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

H. R. Rep. No 111-288 (2009).

Huang, M-H. (2009). Race of the interviewer and the black-white test score gap. *Social Science Research, 38,* 29-38.

Inzlicht, M. & Ben-Zeev, T. (2000). A threatening intellectual environment: Why females are susceptible to experiencing problem-solving deficits in the presence of males. *Psychological Science, 11*, 365-371.

Inzlicht, M. & Schmader, T. (Eds.) (2012). *Stereotype threat: Theory, process, and application*. New York: Oxford University Press.

Jencks, C. (1998). Racial bias in testing. In C. Jencks & M. Phillips (Eds.). *The Black–White test score gap* (pp. 55-85). Washington, DC: Brookings Institution.

Jensen, A.R. (1980). *Bias in mental testing*. New York: Free Press

Johns, J., Inzlicht, M., & Schmader, T. (2008). Stereotype threat and executive resource depletion: Examining the influence of emotion regulation. *Journal of Experimental Psychology: General, 137,* 691-705.

Johns, M., Schmader, T., & Martens, A. (2005). Knowing is half the battle: Teaching stereotype threat as a means of improving women's math performance. *Psychological Science, 16,* 175-179.

Kang, J. & Banaji M. R. (2006). Fair measures: A behavioral realist revision of "affirmative action". *California Law Review, 94,* 1063-1118.

Kobrin, J. L., Sathy, V., & Shaw, E. J. (2007). *A historical view of subgroup performance differences on the SAT reasoning test* (College Board Research Report No. 2006–5). New York: College Board.

Kray, L. J., Thompson, L., & Galinsky, A. (2001). Battle of the sexes: Gender stereotype confirmation and reactance in negotiations. *Journal of Personality and Social Psychology, 80,* 942–958.

Ladson-Billings, G. (2006). From the achievement gap to the education debt: Understanding achievement in U.S. schools. *Education Researcher, 35*, 3-12.

Lerner, M. J. (1980). *The belief in a just world: A fundamental delusion*. New York: Plenum.

Linn, R. L. & Werts, C. E. (1971). Considerations for studies of test bias. *Journal of Educational Measurement, 8,* 1-4.

Logel, C., Iserman, E. C., Davies, P. G., Quinn, D. M., & Spencer, S. J. (2009a). The perils of double consciousness: The role of thought suppression in stereotype threat. *Journal of Experimental Social Psychology, 45,* 299-312.

Logel, C., Walton, G. M., Spencer, S. J., Iserman, E. C., von Hippel, W., & Bell, A. (2009b). Interacting with sexist men triggers social identity threat among female engineers. *Journal of Personality and Social Psychology, 96,* 1089-1103.

Logel, C., Walton, G. M., Peach, J., Spencer, S. J., & Zanna, M. P. (in press). Unleashing latent ability: Implications of creating stereotype-safe environments for college admissions. *Educational Psychologist*.

Martens, A., Johns, M., Greenberg, J., & Schimel, J. (2006). Combating stereotype threat: The effect of self-affirmation on women's intellectual performance. *Journal of Experimental Social Psychology, 42,* 236-243.

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.

Marx, D. M. & Roman, J. S. (2002). Female role models: Protecting women's math test performance. *Personality and Social Psychology Bulletin, 28,* 1183-1193.

Massey, D. S. & Fischer, M. J. (2005). Stereotype threat and academic performance: New findings from a racially diverse sample of college freshmen. *Du Bois Review, 2,* 45-67.

McIntyre, R. B., Paulson, R. M., & Lord, C. G. (2003). Alleviating women's mathematics stereotype threat through salience of group achievements. *Journal of Experimental Social Psychology, 39,* 83-90.

McGlone, M. S. & Aronson, J. (2007). Forewarning and forearming stereotype-threatened students. *Communication Education,* 56, 119-133.

Mendoza-Denton, R., & Page-Gould, E. (2008). Can cross-group friendships influence minority students' well being at historically White universities? *Psychological Science, 19*, 933-939.

Mendoza-Denton, R., Downey, G., Purdie, V. J., Davis, A., & Pietrzak, J. (2002). Sensitivity to status-based rejection: Implications for African American students' college experience. *Journal of Personality and Social Psychology, 83,* 896-918.

Miyake, A., Smith-Kost, L. E., Finkelstein, N. D., Pollock, S. J., Cohen, G. L., & Ito, T. A. (2010). Reducing the gender achievement gap in college science: A classroom study of values affirmation. *Science, 330,* 1234-1237.

Murphy, M. C., Steele, C. M., & Gross, J. J. (2007). Signaling threat: How situational cues affect women in math, science, and engineering settings. *Psychological Science, 18*, 879-885.

Nguyen, H. D. & Ryan, A. M. (2008). Does stereotype threat affect test performance of minorities and women? A meta-analysis of experimental evidence. *Journal of Applied*

Psychology, 93, 1314-1334.

Oreskes, N. & Conway E. M. (2010). *Merchants of doubt: How a handful of scientists obscured the truth on issues from tobacco smoke to global warming*. Bloomsbury Press: New York.

Oswald, F. L., Schmitt, N., Kim, B. H., Ramsay, L. J., & Gillespie, M. A. (2004). Developing a biodata measure and situational judgment inventory as predictors of college student performance. *Journal of Applied Psychology, 89,* 187-207.

Parents Involved in Community Schools v. Seattle School Dist. No. 1, 551, No. 05-908 (2007).

Phillips, M., Brooks-Gunn, J., Duncan, G. J., Klebanov, P., & Crane, J. (1998). Family background, parenting practices, and the Black-White test score gap. In C. Jencks & M. Phillips (Eds.). *The Black–White test score gap* (pp. 103-148). Washington, DC: Brookings Institution.

Pinel, E. C. (1999). Stigma consciousness: The psychological legacy of social stereotypes. *Journal of Personality and Social Psychology, 76*, 114-128.

Primus, R. A. (2003). Equal protection and disparate impact: Round three. *Harvard Law Review, 117*, 494-587.

Purdie-Vaughns, V., Steele, C. M., Davies, P. G., Ditlmann, R., & Crosby, J. R. (2008). Social identity contingencies: How diversity cues signal threat or safety for African Americans in mainstream institutions. *Journal of Personality and Social Psychology, 94,* 615-630.

Reardon, S. F., Atteberry, A., Arshan, N., & Kurlaender, M. (2009). *Effects of the California High School Exit Exam on student persistence, achievement, and graduation* (Working paper). Stanford, CA: Stanford University, Institute for Research on Education Policy and Practice.

Rice v. Cayetano, 528 U. S. 495 (2000).

Roberson, L., & Kulik, C. T. (2007). Stereotype threat at work. *Academy of Management Perspectives, 21*, 24–40.

Rosenthal, H. E. S. & Crisp, R. J. (2006). Reducing stereotype threat by blurring intergroup boundaries. *Personality and Social Psychology Bulletin, 32,* 501-511.

Sackett, P. R., Borneman, M. J., & Connelly, B. S. (2008). High-stakes testing in higher education and employment: Appraising the evidence for validity and fairness. *American Psychologist, 63,* 215-227.

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.

Sackett, P. R. & Ryan, A. M. (2012). Concerns about generalizing stereotype threat research findings to operational high-stakes testing. In M. Inzlicht & T. Schmader (Eds.) *Stereotype threat: Theory, processes, and application*. (pp. 249-263). Oxford, England: Oxford University Press.

Sackett, P. R. & Wilk, S. L. (1994). Within-group norming and other forms of score adjustment in preemployment testing. *American Psychologist, 49,* 929-954.

Sander, R. H. (2004a). A reply to critics. *Stanford Law Review, 57,* 1963-2016.

Sander, R. H. (2004b). A systematic analysis of affirmative action in American law schools. *Stanford Law Review, 57,* 367-483.

Sedlacek, W. E. (2004). *Beyond the big test: Noncognitive assessment in higher education*. San Francisco: Jossey-Bass.

Schmader, T., Johns, M., & Forbes, C. (2008). An integrated process model of stereotype threat

effects on performance. *Psychological Review, 115(2),* 336-356.

Shapiro, J. R. & Neuberg, S. L. (2007). From stereotype threat to stereotype threats: Implications of a multi-threat framework for causes, moderators, mediators, consequences, and interventions. *Personality and Social Psychology Review, 11,* 107-130.

Sherman, D. K., & Cohen, G. L. (2006). The psychology of self-defense: Self-affirmation theory. In M. P. Zanna (Ed.) *Advances in Experimental Social Psychology* (Vol. 38, pp. 183-242). San Diego, CA: Academic Press.

Sherman, D.K. & Hartson, K.A. (2011). Reconciling self-defense with self-criticism: Self-affirmation theory. In M. D. Alicke & C. Sedikides (Eds). *Handbook of Self-Enhancement and Self-Protection* (pp. 128-151). New York: Guilford Press.

Sherman, D. K., Hartson, K. A., Binning, K. R., Purdie-Vaughns, V., Garcia, J., Barba-Taborsky, S., Tomassetti, S., Nussbaum, A. D., & Cohen, G. L. (under review). Self-affirmation, identity threat, and academic underperformance: Understanding the effects of a social psychological intervention.

Society for Industrial and Organizational Psychology (2003). *Principles for the validation and use of personnel selection procedures*. Bowling Green, OH: Society for Industrial and Organizational Psychology.

Son Hing, L. S., Bobocel, D. R., & Zanna, M. P. (2002). Meritocracy and opposition to affirmative action: Making concessions in the face of discrimination. *Journal of Personality and Social Psychology, 83,* 493-509.

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

Steele, C. M. (1997). A threat in the air: How stereotypes shape intellectual identity and

performance. *American Psychologist, 52,* 613-629.

Steele, C. M. (2010). *Whistling Vivaldi: And other clues how stereotypes affect us*. New York: Norton & Company Inc.

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

Steele, C. M., Spencer, S. J., & Aronson, J. (2002). Contending with group image: The psychology of stereotype and social identity threat. In M. P. Zanna (Ed.), *Advances in experimental social psychology, 34*. San Diego, CA: Academic Press.

Steele, C. M., Spencer, S., Nisbett, R., Hummel, M., Harber, K., Schoem, D., & Carter, K. (1998). *African American college achievement: A "wise" intervention*. Unpublished manuscript, Stanford University.

Sternberg, R. J. et al. (2006). The Rainbow Project: Enhancing the SAT through assessments of analytical, practical, and creative skills. *Intelligence, 34,* 321-350.

Sue, D. W., Capodilupo, C. M., Torino, G. C., Bucceri, J. M., Holder, A. M. B., Nadal, K. L., & Esquilin, M. (2007). Racial microaggressions in everyday life: Implications for clinical practice. *American Psychologist, 62,* 271-286.

Stricker, L. J. & Ward, W. C. (2008). Stereotype threat in applied settings re-examined: A reply. *Journal of Applied Social Psychology, 38,* 1656-1663.

Taylor, D. M., Wright, S. C., Moghaddam, F. M., & Lalonde, R. N. (1990). The personal/group discrimination discrepancy: Perceiving my group, but not myself, to be a target for discrimination. *Personality and Social Psychology Bulletin*, *16*, 254-262.

Van den Bergh, L., Denessen, E., Hornstra, L., Voeten, M., & Holland, R. W. (2010). The implicit prejudiced attitudes of teachers: Relations to teacher expectations and the ethnic

achievement gap. *American Educational Research Journal, 47*(2), 497-527.

Walton, G. M. & Carr, P. B. (2012). Social belonging and the motivation and intellectual achievement of negatively stereotyped students. In M. Inzlicht & T. Schmader (Eds.) *Stereotype threat: Theory, processes, and application* (pp. 89-106). Oxford, England: Oxford University Press.

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

Walton, G. M. & Cohen, G. L. (2007). A question of belonging: Race, social fit, and achievement. *Journal of Personality and Social Psychology, 92,* 82-96.

Walton, G. M. & Cohen, G. L. (2011). A brief social-belonging intervention improves academic and health outcomes of minority students. *Science, 331*, 1447-1451.

Walton, G. M., Cohen, G. L., Garcia, J., Apfel, N., & Master A. (2012). *A brief intervention to buttress middle school students' sense of social-belonging:* Manuscript in preparation, Stanford University, Stanford, CA.

Walton, G. M., Logel, C., Peach, J., Spencer, S. J., & Zanna, M. P. (2012). *Two interventions to boost women's achievement in engineering: Social-belonging and self-affirmation-training*. Manuscript in preparation, Stanford University, Stanford, CA.

Walton, G. M. & Spencer, S. J. (2009). Latent ability: Grades and test scores systematically underestimate the intellectual ability of women and ethnic minority students. *Psychological Science, 20,* 1132-1139.

Wax, A. L. (2008). Stereotype threat: A case of overclaim syndrome? U of Penn Law School, Public Law Research Paper No. 08-14, American Enterprise Institute.

Wilson, T. D., Damiani, M., & Shelton, N. (2002). Improving the academic performance of

college students with brief attributional interventions. In J. Aronson (Ed.). *Improving academic achievement: Impact of psychological factors on education*. Oxford, England: Academic Press.

Yeager, D. S. Purdie-Vaughns, V., Hessert, W. T., Williams, M. E., Garcia, J., Apfel, J., Pebley, P., Master, A., & Cohen, G. L. (2011). *Communicating high standards and personal assurance reduces racial achievement gaps*. Manuscript in preparation, Stanford University, Stanford, CA.

Yeager, D. S. & Walton, G. M. (2011). Social-psychological interventions in education: They're not magic. *Review of Educational Research, 81*, 267-301.

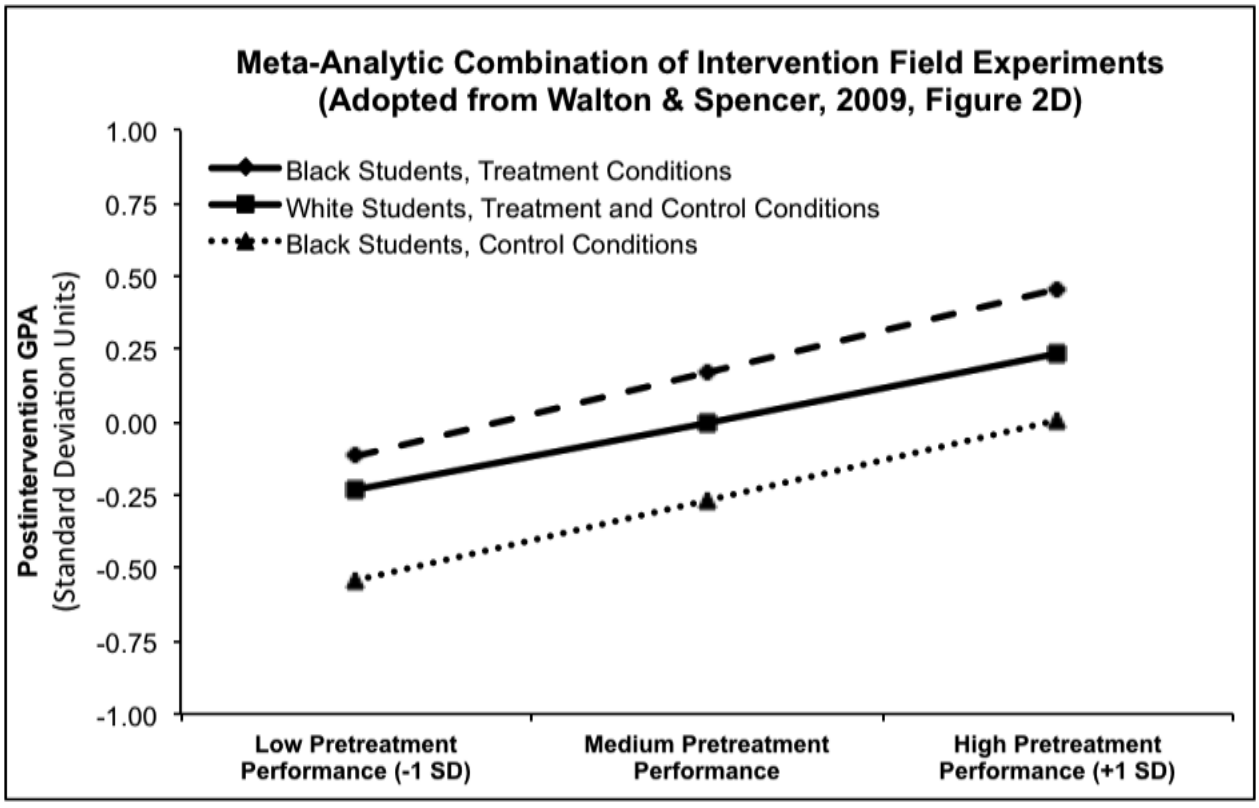
Table 1. Estimated proportion of group differences on the SAT due to stereotype-related psychological threat and to socioeconomic status (SES).

	Estimate of Psychological Threat			
	None	$d = 0.15$	$d = 0.20$	$d = 0.25$
Mean SAT-Math score: Men	531	-	-	-
Mean SAT-Math score: Women/Points lost due to threat	500	18	23	29
Gender gap/% of gender gap due to threat	31	57%	75%	94%
Mean SAT score: Whites	1063	-	-	-
Mean SAT score: Latinos/Points lost due to threat	914	35	46	58
White-Latino gap/% of White-Latino gap due to threat	149	23%	31%	39%
% of White-Latino gap due to SES and threat	46%	70%	77%	85%
Mean SAT score: Blacks/Points lost due to threat	855	35	46	58
White-Black gap/% of White-Black gap due to threat	208	17%	22%	28%
% of White-Black gap due to SES and threat	31%	48%	54%	59%

Note. Means derived from data published by the College Board (2011). SAT is the sum of SAT-Math (SD = 117) and SAT-Reading (SD = 114). Latinos comprise Mexicans/Mexican Americans, Puerto Ricans, and other Hispanics/Latinos/Latin Americans. Analyses assume that SES (i.e., "parental education, family income, and course-taking patterns") explains 49% and 44% of the White-Latino gaps and 30% and 33% of the White-Black gaps on the SAT-Math and -Reading tests, respectively (Camara & Schmidt, 1999, p. 7). These estimates may be conservative. A broader definition of SES may account for more variance (Phillips et al., 1998). Analyses also assume that the effects of SES and psychological threat are independent. But if the contribution of SES to racial/ethnic achievement gaps is mediated in part by stereotype threat—

perhaps low SES minorities perform worse because of stereotypes that target this intersectional identity—then the combination of the two effects may be less than the sum of their main effects. However, we know of no studies that partial out these effects.

Figure 1. *Classroom performance of Black and White students in control conditions and in treatment conditions designed to reduce stereotype-related threat as a function of prior academic performance (grades and standardized test scores). The figure shows the meta-analytic combination of results from the value-affirmation intervention (Cohen et al., 2006), the 21st Century Program (Steele, 1997; Steele et al., 1998), and the social-belonging intervention (Walton & Cohen, 2007, 2011). The latent ability effect is the difference between Black students in treatment conditions and White students at each level of pretreatment performance (*ps < *.005). The meta-analysis of stereotype threat laboratory experiments yielded parallel results (Walton & Spencer, 2009). GPA = grade point average. Adopted from Walton and Spencer (2009), Figure 2D.*



Notes

[1] More research on stereotype threat has examined school performance than work performance. Thus how threat arises, affects performance, and can be reduced is better understood within school contexts. Nonetheless, a consideration of stereotype threat in work settings remains important (see Roberson & Kulik, 2007). For instance, stereotype threat can cause decrements in performance on personnel tests (Gonzales, Blanton, & Williams, 2002) and on important work tasks like negotiations (Kray, Thompson, & Galinsky, 2001) and in non-performance outcomes that affect professional success such as people's sense of belonging in stereotype-laden fields, trust of companies, and career aspirations and motivation (Cheryan, Plaut, Davies, & Steele, 2009; Davies, Spencer, Quinn, & Gerhardstein, 2002; Murphy, Steele, & Gross, 2007; Purdie-Vaughns, Steele, Davies, Ditlmann, & Crosby, 2008). It is thus important to consider both how stereotype threat affects measures of academic performance employers use to make hiring decisions and how threat can be reduced within work contexts.

[2] A methodological question involves the comparison group used to test latent ability. The meta-analysis of laboratory experiments compared stereotyped students in nonthreatening conditions to nonstereotyped students in nonthreatening conditions. This is because nonstereotyped students experience a performance boost called *stereotype lift* when they are aware that an outgroup is negatively stereotyped in a testing situation (Walton & Cohen, 2003). Comparing stereotyped students to nonstereotyped students in threat conditions would thus introduce a benefit for one group and not the other and so create a confound. In the subsequent meta-analysis of intervention field experiments, White students' performance did not vary by condition so, to maximize statistical power, this meta-analysis compared stereotyped students in treatment conditions to nonstereotyped students in both conditions.

Sackett & Ryan (2012) reject this comparison in the meta-analysis of laboratory experiments. They write that the estimate of nonstereotyped students' ability is "downwardly biased" (p. 255) in nonthreatening conditions because in this condition they do not benefit from stereotype lift. Instead, they advocate comparing stereotyped students (in nonthreatening conditions) to nonstereotyped students in threatening conditions. Yet if the best estimate of ability is one that occurs when people benefit from a negative stereotype about an outgroup, then performance in nonthreatening conditions—which take stereotypes off the table for everyone—is "downwardly biased" for both groups. From this perspective, in the ideal comparison both groups would benefit from stereotype lift; such data do not exist. At least in the present comparison the ostensible "bias" extends to both groups. Sackett and Ryan's preferred comparison would, without justification, estimate nonstereotyped students' ability while they benefit from stereotype lift and stereotyped students' ability without this benefit. Additionally, this critique simply ignores the parallel evidence for latent ability from the meta-analysis of intervention field experiments. Not only does this second meta-analysis replicate the first meta-analysis; the results from the latter are, if anything, more pertinent as it estimates the degree to which common performance measures underestimate the academic potential, growth, and learning of students in a new environment. To the extent that this is exactly what college admissions tests like the SAT are intended to do, this bias is highly problematic and requires remedy.

[3] Two technical reasons also suggest that the observed magnitude of the latent ability effect underestimates the bias (see Linn & Werts, 1971). Both reasons issue from the possibility that stereotyped groups actually have, on average, less developed academic ability than nonstereotyped groups (e.g., as a result of prior disadvantage). If so, regardless of how well they

perform on an initial measure, members of stereotyped groups will tend to regress to a lower group mean on a subsequent measure than members of nonstereotyped groups; they will thus perform less well at every level of prior performance. This is a consequence of imperfect reliability in performance measures, and will occur even if performance measures are nonbiased. The same effect will occur if the predictor (e.g., SAT scores) assesses only a portion of the variance in students' ability, if group differences exist in unmeasured aspects of ability, and if those unmeasured aspects of ability contribute to the criterion (e.g., college grades). These factors cause interpretive ambiguity when underperformance is observed. But when latent ability is observed, they lend confidence to the effect and suggest that the observed effect sizes underestimate the bias.

[4] A reader may ask whether threat-reducing strategies could harm nonstereotyped students by reducing their benefit from stereotype lift, the performance boost nonstereotyped students experience when they are aware that a negative stereotype targets an outgroup (Walton & Cohen, 2003). Contrary to this supposition, interventions to reduce stereotype threat in field settings either do not affect nonstereotyped students (e.g., value-affirmation and social-belonging interventions, Cohen et al., 2006; Walton & Cohen, 2011) or benefit nonstereotyped students (e.g., theory of intelligence interventions, Aronson, Fried, & Good, 2002). This may be because these interventions mitigate the effects of stereotypes only indirectly, reducing their impact on stereotype lift, and because they include elements that benefit all students. However, one study found a negative effect for nonstereotyped students. Moving demographic queries from before to after the AP calculus test seemed to lower boy's scores (Danaher & Crandall, 2008), an effect that is consistent with a reduction in stereotype lift, although the benefit of this intervention to girls exceeded its effect among boys. When such effects are found, complex questions about the

interpretation of performance measures arise. To the extent that we aspire to create settings in which social stereotypes do not impinge upon performance—to create, for instance, stereotype-safe college math classes and professional settings—our view is that performance is best assessed when group stereotypes are off the table for everyone.

[5] An important question is whether organizations that observe *under*performance should correct scores in the opposite direction—disadvantaging stereotyped students. We think not. Sackett and Wilk (1994) note that such score corrections would have negative "social and political consequences" (p. 933). For instance, they would be morally problematic—they would punish talented stereotyped students for an organization's failure to reduce threat.

[6] As noted, courts closely scrutinize actions taken by public entities and recipients of federal funds that differentiate people on the basis of race or sex (Title VI of the Civil Rights Act of 1964; Constitution). Here we focus on racial classifications. The Supreme Court recognizes promoting diversity as an interest that can justify limited racial classifications. It permits higher-education organizations to promote diversity through individualized admissions decisions that take race into account. But it prohibits organizations from using mechanical, race-based allocations of points (*Gratz v. Bollinger*, 2003; *Grutter v. Bollinger*, 2003).

Racial classifications in affirmative meritocracy serve a different end—to promote merit. As the Court has not addressed merit systematically it is not known whether it would recognize merit as an interest that can justify differentiating individuals based on race. But its reasoning is protective of merit. The Court has criticized racial classifications as anti-meritocratic, writing, "it demeans the dignity and worth of a person to be judged by ancestry instead of by … merit" (*Rice v. Cayetano*, 2000, p. 517, quoted in *Parents Involved in Community Schools v. Seattle School Dist. No. 1*, 2007, p. 39 (plurality opinion)). It has sought to curb the use of racial classifications

by prohibiting schools from "insulat[ing] [candidates] … from competition" with one another (*Grutter v. Bollinger*, 2003, p. 334). In rare cases when the Court recognizes that the absence of racial classifications would undermine merit (e.g., if a school chose to promote diversity by admitting students using a lottery) it instead allows schools to promote diversity using limited racial classifications (*Grutter v. Bollinger*, 2003, p. 334). The Court's solicitude for merit suggests that merit itself could justify limited racial classifications. If so, organizations could likely implement an individualized selection process like that described in the main text. Further, unlike the similar process the Court upheld in *Grutter*, where the Court evinced concern that affirmative action had no end point, the racial classifications of affirmative meritocracy are, as noted, self-extinguishing.