Stereotype Threat and Employment Testing: A Commentary

Claude M. Steele and Paul G. Davies

Department of Psychology
Stanford University

This article comments on Sackett, Schmitt, Ellingson, and Kabin’s (2001) suggestion that stereotype threat may not generalize to employment testing, and the 4 articles of this issue following from it. We argue that each experiment lacked the no-stereotype threat control group needed to experimentally test stereotype threat effects; that sizeable stereotype threat effects occurred nonetheless; and that research not reviewed by Sackett et al. suggests that stereotype threat may be the sole cause of important real-life test-score gaps. The 4 experiments, we suggest, reveal important mediators of stereotype threat effects; their test-specific and general form, and their generalizability to IQ testing. We offer a framework for assessing its generalizability to real-life employment tests.

The studies included in this special issue were motivated by the important question of generalizability, in particular, whether stereotype threat can affect performance on real-life employment tests. Stereotype threat is the pressure that a person can feel when she is at risk of confirming, or being seen to confirm a negative stereotype about her group. Of greatest relevance here are the forms of this threat that stem from negative stereotypes about a group’s intellectual abilities. This form of stereotype threat has been shown to impair the intellectual test performance of multiple groups (women, African Americans, Latinos, French students from lower socioeconomic status [SES] backgrounds, White men, etc.) on multiple tests (sections of the GRE, SAT, GMAT, LSAT, etc.; cf., Steele, Spencer, & Aronson, 2002). However, Sackett, Schmitt, Ellington, and Kabin (2001) argued that although early stereotype threat studies show that stereotype threat manipulations can “expand” the conventional Black–White test-score gap in laboratory studies, they do not show that stereotype
threat contributes to this gap in real-world employment testing. The four articles in this issue address this generalizability question, among others. In responding to this question and to these articles, we will make a general point.

As noted, an impairing effect of stereotype threat on cognitive test performance has been shown in multiple groups, on many tests, in laboratory and real-life testing situations, in many countries, by many different investigators, in published experiments that now approach 100 in number. This special issue presents four laboratory studies that report no effect of their experimental manipulations stereotype threat. Null findings for a standard effect are notoriously difficult to interpret—so many causes are possible—especially when no particular cause is specified and experimentally tested. But in this comment we do raise one possibility that we hope will be useful to further research: That following the Sackett et al. (2001) interpretation, the studies included in all four articles have a design limitation—the absence of adequate no-stereotype-threat control groups. This limitation prevents them from providing the experimental test of the stereotype threat hypothesis that they set out to provide.

Beyond this general point we will also make several more specific ones:

1. That the data from these studies can still be interpreted as showing sizeable effects of stereotype threat on test performance.
2. That although offering interesting evidence as to how stereotype threat affects test performance, these studies do not generally address the question of generalizability to real-life employment testing.
3. That the studies reported provide suggestive evidence of: (a) individual differences and other factors that may mediate the effects of stereotype threat; (b) evidence that stereotype threat can take a general form (a stereotype wariness across settings) and a specific form (a stereotype wariness focused on the testing situation) and that these forms can affect test performance under different conditions; and (c) evidence that the detrimental effects of stereotype threat can generalize to IQ tests.
4. We conclude our overview remarks by stressing the contextual nature of stereotype threat—that it is rooted in intersubjectivity, the perception of when a negative stereotype about one’s social identity is relevant to one’s behavior, and thus, when one might be judged by it. Thus the strength of this threat on real-life tests will depend, in part, on the number and strength of cues in the setting that signal the risk of stereotyping, that signal the predicament of stereotype threat. Individual differences and testing instructions that affect the interpretation of those cues will also matter. This view, we suggest, provides a framework for understanding the generalizability of stereotype threat effects to real-life settings, for understanding its variation in strength across settings, and for understanding how to reduce it in those settings.
5. Finally, we return to the Sackett et al. (2001) interpretation. We describe research suggesting that stereotype threat may be the sole cause of important real-life test-score gaps, and we respond to each of the four articles separately.

OVERVIEW COMMENTARY

The Sackett et al. (2001) Interpretation

Some stereotype threat experiments use the analysis of covariance to statistically equate groups on the skills involved in cognitive test performance. For example, Steele and Aronson (1995) used participants’ SAT scores to adjust their scores on a section of the GRE taken in the experiment. The adjusted GRE scores are presumably corrected for the typical difference between these groups in the skills measured by the SAT.

Based on its use of this analysis, Sackett et al. (2001) questioned whether stereotype threat research generalizes to real-life employment testing. Their argument has three parts:

1. Tests like the SAT and GRE measure the same kinds of skills and knowledge. Thus adjusting the in-the-experiment performance of two groups on the GRE for their real-life performance on the SAT should virtually eliminate any group difference on the (adjusted) GRE.
2. If some experimental condition does produce a group difference on the adjusted GRE (e.g., Whites still outperform Blacks), then that condition must have done something to expand the normal performance difference between these groups (because, presumably, the analysis of covariance [ANCOVA] has already equated them for that normal difference).
3. What expands the normal performance difference between these groups in the experiment may not be a factor in real-life employment testing. After all, the ANCOVA presumably equated the groups for their real-life differences in test-relevant cognitive skills. Thus, anything that produced a further difference in the laboratory may not be an actual factor in real-life testing.

This is the Sackett et al. (2001) interpretation as we understand it. They focus it on the first published stereotype threat studies of Steele and Aronson (1995). These studies show that in conditions that have eliminated stereotype threat, Blacks and Whites perform the same on a section of the GRE once their scores were adjusted for their SAT scores. But most importantly, in the stereotype threat
conditions of these experiments, Blacks score almost a full standard deviation worse than Whites even after the SAT adjustment. Following the previous logic, they argued that the laboratory manipulation of stereotype threat must have done something extra that expanded the typical Black–White gap observed in the real world—because participants’ performance was already statistically corrected for this gap. And this something—because it caused a gap that went beyond the typical Black–White gap observed in real-life—may not be something that is typical of real-life employment testing. Stereotype threat, they suggest, may be less a pressure that affects real-life testing, than a pressure that, in the lab, can expand the typical race gap in test performance.

At this point we would like to argue two things at once: (a) that the Sackett et al. (2001) interpretation is not an adequate account of stereotype threat effects, and (b) that it has led each of the four experiments reported in this journal issue to have an important design limitation, one that prevents them from providing an adequate experimental test of the stereotype threat effect. We will return to a critique of Sackett et al. But since our commentary is focused on the experiments reported in this issue, we begin with a description of how Sackett et al. seems to have affected the design of these experiments.

An Unchecked Implication

Following the logic just outlined, Sackett et al. (2001) assumed that the stereotype threat conditions of Steele and Aronson (1995) did something extra—something not representative of real-life testing—to expand the normal Black–White test score gap. Importantly, the no-stereotype threat conditions in these experiments did not expand the Black–White gap once performance was adjusted for SATs. Thus these conditions produced the results that Sackett et al. (2001) said should be expected from use of the covariate. They state “Absent stereotype threat, the Black–White difference [in the Steele & Aronson experiments] was just what one would expect (i.e. zero), as the two groups had been equated on the basis of SAT scores” (p. 31).

Reasonable enough up to this point. However, embedded in this analysis is an implication: If the stereotype threat condition expanded the normal Black–White gap by doing something untypical of real-life testing, then the no-stereotype threat condition must have avoided expanding the Black–White gap by doing what was typical of real-life testing. In this way, the no-stereotype threat condition became the condition that Sackett et al. (2001) saw as typical of real-life testing.

An Implication Begets a Design Limitation

As noted, the researchers in this issue seemed to accept the Sackett et al. (2001) interpretation as a working assumption. Following the previous implication of Sackett et al., they assumed that a no-stereotype threat control condition could be
operationalized by using testing instructions roughly like real-life testing instructions. And they assumed that a stereotype threat condition had to implement an extra degree of threat that went beyond the stereotype threat inherent in conventional real-life testing.

They seemed to miss the actual condition instructions used in Steele and Aronson (1995) and in many subsequently published stereotype threat studies. In Steele and Aronson the no-stereotype condition does not present the test as a test at all, it presents it as an instrument for studying problem solving per se, an instrument that was explicitly described as non-diagnostic of individual differences in ability. The goal was to make the negative racial stereotype irrelevant to Black participants’ performance on this task—and thus, to reduce their felt stereotype threat. Representing the task as not a test but as an instrument for studying problem solving made the negative stereotype about their group’s ability irrelevant to their performance on the task. This is a central feature of the no-stereotype threat control condition in the Steele and Aronson experiments and many subsequent stereotype threat experiments. Finding a way to eliminate stereotype threat in a testing situation is the difficult thing to do. Yet it is necessary to enable a true experimental test of the effect.

It was the stereotype threat conditions themselves that were designed to be like the conditions of real-life testing. These conditions either made it explicit, or allowed participants to assume that they were taking a real test, a test that like most such tests was diagnostic of a cognitive ability and thus was relevant to the negative stereotype about their group.

Following the Sackett et al. implication, rather than the procedures in these stereotype threat experiments, all of the experiments in this issue set up the following comparison to test the effect of stereotype threat: they compared a control condition that presented the test as a test, and an experimental condition that also presented the test as a test, but that added an extra degree of stereotype threat by, say, having the participants list their race before taking the test.

**Overarching Critique of the Four Experiments in This Issue**

In our view, this design feature renders these experiments unable to provide an adequate experimental test of the effect of stereotype threat on employment testing. To test this effect they compare a condition that has aroused stereotype threat with a condition that aroused stereotype threat plus, possibly, a little more stereotype threat. They are not comparing a condition that aroused no stereotype threat with a condition that aroused some stereotype threat. Both the experimental and control conditions in these experiments use procedures that have been shown to produce enough stereotype threat to impair cognitive test performance (cf. Steele et al., 2002).
The Generalizability Question

These studies also do not adequately address the generalizability of stereotype threat effects to real-life employment testing. The most obvious way to do this would be to somehow test the effect of stereotype threat in real-life testing situations. Each of the four studies in this issue present the generalizability question as a primary rationale. Yet none of them were conducted in real-life testing situations. They are laboratory experiments.

The claim that they are addressing the generalizability question seems to rest on the Sackett et al. (2001) interpretation described earlier: that the no-stereotype threat control conditions in these experiments were essentially like the conditions of normal employment testing, and that the stereotype threat conditions involved some extra degree of threat not typical of real-life testing. Then the reasoning must go: because the control conditions were like the conditions of real-life testing, these experiments could test stereotype threat’s generalizability to real-life by testing whether a condition that involved real-life testing procedures plus an explicit manipulation of stereotype threat would impair test performance. If it would, then the researchers could conclude that stereotype threat—as a threat extra to the normal conditions of testing—might well generalize to real-life. If it didn’t, as it didn’t, then they would conclude that it would likely not generalize to real-life since it didn’t cause any impairment beyond that caused by conditions that simulated real-life conditions.

Such is our best rendition of the reasoning involved. But whatever the reasoning, it would seem to encounter two problems. First, as noted earlier, these are laboratory experiments. Their control conditions may reasonably simulate real-life testing procedures. But whatever condition comparisons are set up between the conditions of these experiments, they cannot test the real-world generalizability of an effect because they are lab comparisons not comparisons in the real-world. The second problem is the design limitation raised earlier: in our view, these comparisons do not adequately test the effect of stereotype threat because both conditions in the comparison—the stereotype threat and control conditions—evoke stereotype threat for Black participants. Thus the critical comparison testing stereotype threat is inadequate, and even if it were adequate, it is an in-the-lab comparison. For either reason, the results of this comparison would not say much about the generalizability of stereotype threat effects to real-life.

Evidence of Stereotype Threat Effects in the Four Experiments

But—and this is a big point—there is evidence of a substantial effect of stereotype threat on employment-type tests throughout these experiments. In several of them (Nguyen, O’Neal, & Ryan, this issue, and McFarland, Lev-Arey, & Ziegert, this issue) Blacks scored significantly lower than Whites on an employment-type
test even after group differences in skills had been controlled for, that is, even after their performance had been adjusted for group differences in SAT scores. If Black and White scores were equated for the group difference in SAT, then what could have made Blacks’ adjusted scores in these experiments so much worse than those of Whites?

Our guess, of course, is stereotype threat. As we have surmised, both the stereotype threat and control conditions in these experiments caused stereotype threat in Black participants. That Blacks then scored worse than Whites even after the SAT adjustment, suggests that this surmise is correct. It is precisely this pattern of data that Steele and Aronson (1995) and other stereotype threat studies have used to document this effect. And in the experiments that show these effects, the size of the effects are substantial, similar in magnitude to stereotype threat effects reported in the broader stereotype threat literature.

To definitively claim these effects as stereotype threat effects, one would need more evidence—evidence of their being mediated by felt stereotype threat (the Mayer & Hanges [this issue] study does offer evidence of this sort) and most importantly, experimental evidence showing that the effect does not emerge in a real control condition that arouses no stereotype threat. But our bet is that these group performance effects are caused by stereotype threat. We would bet that this is true as well for the two other experiments in this set (Mayer & Hanges, this issue, and Ployhart, Ziegert, & McFarland, this issue). They also show substantial race effects on performance, but do not appear to statistically control for participants’ SATs. Implementing these controls would likely reduce the size of these effects, but we doubt that it would eliminate them altogether. None of the researchers in this issue note this interpretation of these effects.

**You do not Need to Call a Test a Test to Get Stereotype Threat Effects**

Does a test have to be called a test for test-takers to experience stereotype threat? This question is relevant to the issue of how well stereotype threat effects observed in the laboratory generalize to the conditions of real-life employment testing. Ployhart et al. (this issue), for example, suggested that the procedures used to induce stereotype threat in the laboratory were a bit heavy-handed compared to testing instructions in real-life. They stated “many previous studies have used very strong manipulations of diagnosticity such as telling participants that the test is diagnostic of a particular construct and that the stereotyped group performs more poorly on it.” For reasons specific to particular experiments, these “strong manipulations” have been used in several published studies (e.g., Aronson et al., 1999; Leyens, Desert, Croizet, & Darcis, 2000). But these studies are the exception to the rule. In most stereotype threat studies, the manipulation merely mentions that the test is a measure of ability.
Most importantly, stereotype threat effects have been shown consistently in conditions that say nothing about the test, that simply allow participants to assume that the task is a test (e.g., Davies, Spencer, Quinn, & Gerhardstein, 2002; Spencer, Steele, & Quinn, 1999). This is easy to do of course. Most students simply assume that a GRE test booklet is in fact a GRE test booklet, a test that is designed to measure some kind of cognitive ability. You do not have to be explicit to get them to think this. The difficulty is to get participants in the no-stereotype threat conditions to accept that such a booklet is not a test of cognitive ability.

However, the possibility that stereotype threat effects might not generalize to real-life employment testing because only heavy-handed procedures produce such effects is not a possibility that fits the evidence. Simply having a test that is recognizable as a test is enough to meet the minimal conditions for arousing stereotype threat in the laboratory, and we presume in real-life too.

The Stereotype Threat Literature Itself Should Abate Concerns About Generalizability

Both Sackett et al. (2001) and the articles in this issue seem to over-focus on the early Steele and Aronson (1995) experiments to the neglect of the larger literature on stereotype threat. The effect has now been demonstrated in different groups, on different tests and behaviors, under different conditions, in several countries, by many different investigators. There are nearly 100 published studies showing the effect. Thus the question of whether this effect generalizes to real-life employment testing—the question that motivates the research in this issue—is not compelling to us on the face of it. It would have seemed more compelling several years ago when less stereotype threat research had been completed. But now, the question seems to reflect the unfortunate segregation of these two literatures, the literature on employment testing and the social psychological literature on stereotype threat (see Ployhart et al., this issue). We believe the more apt question is not whether stereotype threat generalizes to real-life employment testing—we do not doubt that it does—but under what conditions.

When Does Stereotype Threat Affect Employment Testing?

The critical thing to keep in mind about stereotype threat is that it is a predicament—the situation of being at risk of confirming or being seen to confirm a negative group stereotype that, if confirmed, could obstruct something the person cares about. Research on stereotype threat shows that this predicament can be a part of the test-taking experience. Thus the impact of this threat on test performance should depend on factors that affect the strength of this predicament—the strength of one’s identification with the domain to which the test is relevant; the difficulty of the test; the strength of one’s identification with the stereotyped group; the stakes
attached to the test; individual differences in strategies for coping with stress and evaluative pressure; personal theories of intelligence; and so on. The more these factors conspire to build a sense of threat around the test-taking experience, the stronger stereotype threat should be.

Thinking this way, we would not assume that stereotype threat affects all employment testing. If the test has many parts that test-takers are familiar with and can succeed on, they may not experience the frustration needed to arouse a concern about how they are doing—and thus not experience a strong predicament of stereotype threat. If the stakes associated with the test are fairly low—for example, if the test is used to assign them to one of several fairly attractive work stations—the test may not cause much stereotype threat. Or if one is not very identified with the domain of material being tested, one may not experience much stereotype threat.

In general, the effect of stereotype threat on real employment tests should be reduced by factors that reduce the predicament posed by the negative stereotype, that is, factors that make one’s performance less likely to confirm the stereotype (e.g., a test focused less at the frontier of knowledge and skills); factors that make confirming it less consequential (e.g., a lower stakes test); or factors that give the test-taker some means of coping with a stereotype confirming performance (e.g., providing alternative routes to success in the domain). And likewise, factors that intensify this predicament should increase the effect of stereotype threat on real-life employment testing. For example, more difficult, timed tests with high stakes attached to them should produce greater performance gaps between groups whose relevant abilities are negatively stereotyped and those whose relevant abilities are not negatively stereotyped. (A recent article [Walton & Cohen, 2003] suggests, in fact, that the non-stereotyped group may actually experience a “stereotype lift” in performance from the conditions that produce stereotype threat in the negatively stereotyped group.)

We have long felt that laboratory studies of stereotype threat may underestimate the size of stereotype threat effects in many real-life testing situations. This is because the laboratory is limited in what it can do to strengthen the predicament of stereotype threat. It cannot do much to intensify the stakes attached to a test performance, for example. Most investigators do what they can to intensify this predicament. They recruit highly identified participants, they use difficult tests, they impose strict time limits on performance, and so on. However, in real-life testing, these conditions are often involved along with very high stakes, something difficult to mimic in the lab—the attempt of these experiments to do so notwithstanding.

Thus, we do not see the question of stereotype threat’s generalizability to employment testing as a question with a yes-or-no answer. In our view, there is little doubt that it generalizes under some conditions—and probably under some important conditions—and not at all under other conditions. For us, the question for future research is in what forms of employment testing are these effects most significant and what can be done to reduce them? In directing this research, we believe
that a particular idea is important to keep in mind: that test-taking is an experience that, especially for groups whose abilities are negatively stereotyped, can involve a powerful predicament of identity, and that to understand and improve their performance, one must understand and address this predicament.

Returning to the Sackett et al. (2001) Interpretation

We would like to make two general points about Sackett et al. (2001): one about the significance of literature they have overlooked, and the other about an assumption they make.

A point about evidence. The comments of Sackett et al. (2001) were apparently provoked by media reports implying that the entire Black–White test-score gap was attributable to stereotype threat. These reports, they noted, overlooked the significance of equating Blacks and Whites in the Steele and Aronson (1995) experiments on their SATs—a significance spelled out earlier. What they wanted the public to understand is that this gap had more causes than just stereotype threat, and that it would likely persist even if stereotype threat were eliminated. We have long shared this concern about overstatements of the role of stereotype threat in the Black–White test-score gap. We have always stressed that other factors, such as racial inequalities in educational opportunity, have to play a role in this gap.

But we would be remiss if we did not note that important stereotype threat effects have emerged in the literature that do not involve any use of covariates. These effects suggest that some very important group differences in test performance may be caused entirely by stereotype threat.

For example, Croizet and Claire (1998) showed that stereotypes about social class and intelligence had a strong effect on the test performance of French college students. They found that lower SES French students underperformed in relation to high SES students on a GRE-type test when participants were allowed to assume the test was diagnostic of ability—a condition that put lower SES test-takers under stereotype threat, that is, under the threat of confirming, or being seen to confirm the negative stereotype about their group’s ability. However, when the test was represented as a non-diagnostic task—making the negative class stereotype irrelevant to the performance of the low SES students—the lower SES students performed as well as the high SES students. In these experiments, no prior test score was used as a covariate to equate the SES groups in skills. This means that the underperformance of the low SES students under normal testing instructions (the diagnostic conditions) was caused entirely by stereotype threat. These results have now been replicated many times, and have been extended to other tests such as the Raven’s Advanced Progressive Matrices (APM) Test. The same has been shown in many experiments examining the effect of stereotype threat on women’s math test performance (e.g., Davies, Spencer, Quinn, & Gerhardstein, 2002).
Now, of course, participants in these studies are all college students. They are taken from selected populations and thus do not include the full range of skills that exist within these groups in the larger population.

Still, these studies can be safely assumed to generalize to at least the college population. And, they did not use a covariate. Thus none of the arguments raised by Sackett et al. (2001) about the real-life generalizability of Steele and Aronson (1995) would apply to them. Moreover, as noted, we assume that the laboratory context underestimates the strength of the stereotype threat pressures that bear on real-life testing. Thus the findings of these experiments carry a profound implication: that the entire underperformance of these important college groups—women in math and lower SES test-takers more generally—may be due to stereotype threat, and that the size of this underperformance in real-life may be even greater than it is in these experiments.

A point about assumptions. Now we have in place the reasoning we need to address a fundamental assumption of the Sackett et al. (2001) interpretation. Recall their analysis of the Steele and Aronson (1995) experiments: that adjusting participants’ GRE scores in the experiment for their SAT scores must have equated Blacks and Whites on the factors that affect test performance in real-life testing. Thus, if the stereotype threat condition caused Blacks to perform lower than Whites despite this adjustment, it must have caused an extra pressure, a pressure not typical of real-life testing.

We might agree that the stereotype threat condition put an extra pressure on Black participants’ GRE performance, one that was not present or was weaker for their SATs. The GRE exam could presumably prognosticate their future experience with such tests, they were highly identified with their performance, and it was an extremely frustrating test, pre-tested so that participants would get no more than 25% of the items correct on average. And for the Black students, they were taking this test in an atmosphere that would likely accentuate stereotype threat compared to their high school environments—on a very selective campus in implicit competition with very strong students. The SAT experience for these highly selected students may have been a lot easier, and they may not have experienced much stereotype threat taking that exam.

However, there is no reason to assume that the pressures bearing on the GRE are not generalizable to real-life testing. There are clearly employment tests that are difficult and that are taken by people who are identified with succeeding on them. That is all that would have to be similar for these results to generalize to such testing situations. Stereotype threat is not an absolute thing that either generalizes or not. As explained earlier, it is a predicament that varies in strength depending on the specific elements of the testing situation. As long as there are real-life employment testing situations that involve the same elements as the stereotype threat conditions in the Steele and Aronson (1995) experiments, then the results of those
experiments—their demonstration of the performance-impairing effects of stereotype threat—should generalize to those real-life testing situations. We believe that this is how to think about generalizability.

COMMENTARY ON THE FOUR ARTICLES IN THIS ISSUE

McFarland, Lev-Arey, and Ziegert

McFarland et al. examined whether stereotype threat effects would occur among participants who were more highly motivated to perform, and whether these effects would generalize to stereotype-irrelevant tests administered during the same testing session. To examine these issues the researchers used a 2 (race: Black, White) × 2 (condition: threat, no-threat) × 2 (order of administration: personality test first, intelligence test first) between-subjects design. They also created a higher-stakes testing scenario. This is an excellent idea. As the researchers correctly note, very little stereotype threat research has been conducted in paradigms that involve an explicit motivational incentive. McFarland et al. reason that stereotype threat effects may be weakened in higher-stakes situations. That is, with stronger motivation, stereotype susceptible participants may push through this pressure to perform well. As we have long stressed, we would expect the opposite; that increased motivation would only intensify susceptibility to stereotype threat. Regardless of the particular hypothesis, however, the idea of increasing motivation in the stereotype threat paradigm is long overdue.

McFarland et al. informed all of their participants, both verbally and in writing, that the study was being conducted to examine test performance on two types of measures. To create a higher-stakes testing scenario, participants were also told that those who performed in the top 15% across both tests would receive a $20 bonus. Participants in the stereotype threat condition were then asked to complete a racial identity scale and a demographic form, along with being informed that the tests they were about to take were good indicators of intelligence and personality. Participants in the no stereotype threat condition were simply told they would be taking a personality test and a problem-solving task. These instructions to participants in the no-threat condition would not likely elicit stereotype threat among Black participants on their own. But when these problem-solving instructions are heard within the context of a study examining test performance that rewards only “those who achieve scores in the top 15%” it is not unlikely that they would elicit stereotype threat. Similar instructions have been sufficient to do so in other stereotype threat research (cf. Steele et al., 2002).

In fact, MacFarland et al.’s findings for intelligence test performance fit this interpretation. Even after Blacks and Whites had been equated on SAT scores,
Blacks performed worse than Whites in both threat conditions. It is at least possible that Blacks experienced stereotype threat in both conditions. The experiment did not find the predicted interaction between race and condition, or a main effect for condition. Regardless of condition, Black participants consistently underperformed.

Thus, the principal challenge for this research paradigm is to create a condition within a high-stakes testing environment in which Black participants are not susceptible to stereotype threat, that is, to create an effective no-stereotype control condition.

Turning to the personality test data, McFarland et al. reported that only the main effect for condition was significant, meaning that both Black and White participants in the stereotype threat condition performed better on the personality test. As the authors note, this is not a test that is likely to evoke stereotype threat. Thus the failure of race to affect performance on this test, coupled with its effect on the stereotype-threat evoking intelligence test, fits the view that Blacks’ underperformance on the intelligence test was caused by stereotype threat.

Mayer and Hanges

The Mayer and Hanges article expands on previous research conducted by McKay, Doverspike, Bowen-Hilton, and Martin (2002), which examined the effect of stereotype threat on the Raven APM test. They used a 2 (race: Black, White) × 2 (test diagnosticity: intelligence, perceptual ability) between-subjects design. These researchers did a commendable job of creating a testing environment that was more reflective of those found in actual personnel testing situations. As importantly, they explored possible mediators of stereotype threat effects, with a special focus on two separate forms of stereotype threat (i.e., specific and general). Their paradigm has a great deal of potential for future research on stereotype threat.

This experiment, however, may not have manipulated test diagnosticity—and thus stereotype threat—successfully. As the authors stated, “contrary to expectations, participants were no more likely to believe that the Raven’s APM test measured intelligence if they were in the diagnostic as opposed to the non-diagnostic condition.” Without a successful diagnosticity manipulation it is impossible to have a successful stereotype threat manipulation. This could explain why Mayer and Hanges’ did not replicate the McKay et al. (2002) finding that Blacks underperformed more in relation to Whites when the Raven APM test was framed as diagnostic of intelligence.

However, Mayer and Hanges did find that Blacks underperformed in relation to Whites in both diagnosticity conditions. Again, it is conceivable that both conditions caused enough stereotype threat to undermine the performance of Black participants. What is needed to answer this question more definitively is a control condition that successfully eliminates stereotype threat for the Black participants.
The diagnosticity manipulation in this experiment may have been undermined by one of the six self-report measures (i.e., anxiety, cognitive interference, self-efficacy, evaluation apprehension, predictive validity, and belief in tests) that participants completed after the manipulation and before the Raven APM test. For example, Mayer and Hanges reported that one of the items from the predictive validity questionnaire reads: “I am confident that this examination can predict how intelligent a person is.” That item alone could be enough to counteract non-diagnostic instructions.

Finally, the authors’ effort to distinguish between two forms of stereotype threat (i.e., specific and general) offered interesting and promising findings. Both stereotype threat-specific (a stereotype wariness focused on the testing situation) and stereotype threat-general (a stereotype wariness across settings) were negatively related to Raven APM test performance. This is an important finding of both theoretical and practical significance. We hope the authors continue to investigate how these two forms of stereotype threat affect test performance across groups, situations and tests.

Ployhart, Ziegert, and McFarland

Ployhart et al. maintained that Blacks perceive cognitive ability tests as being less valid than do Whites, which may influence test-taking motivation and performance. They reason that face validity manipulations that improve test perceptions among Blacks may consequently improve their test performance. To examine these issues, Ployhart et al. designed a study that simulated a selection context for a retail manager, in which both stereotype threat and the face validity of the selection test were manipulated. They used a 3 (stereotype threat: diagnostic, nondiagnostic, control) × 2 (face validity: face valid, generic) × 2 (race: Black, White) between-subjects design. All participants were informed that the top 15% of performers on the selection test for retail managers would receive a $20 bonus. In the diagnostic condition the test was described as “designed to measure your intelligence.” In the nondiagnostic condition it was described as “designed to measure your skills as a retail manager,” and in the control condition it was simply described as a “difficult test.”

The manipulation of the test’s face validity was creative. For the Generic condition the test was labeled the “General Intelligence and Aptitude Test” and was composed of 21 GRE and GMAT sample questions. For the Face Valid condition the test was labeled the “Retail Management Skills Test,” and was composed of the same 21 questions rewritten to relate directly to a retail manager position. Ployhart et al. pretested the two versions of the test asking participants the degree to which the items measured intelligence and retail manager potential. They report that participants viewed the retail items as assessing retail manager potential more than the generic items. But they did not report whether the generic items were viewed as as-
sessing intelligence more than the retail items. This essential manipulation check is missing from the article.

It is needed to help interpret several findings. As in the other experiments reported in this issue, this study too found that Black participants underperformed in relation to White participants across conditions that were designed to vary stereotype threat—suggesting that the procedures in all of these condition put Blacks under stereotype threat. Also, although underperforming in relation to Whites, Blacks themselves performed about as badly regardless of whether they were in a condition in which the test was labeled General Intelligence and Aptitude Test or Retail Management Skills Test. Clearly, one would expect the General Intelligence and Aptitude Test condition to arouse stereotype threat and undermine performance. Thus, for the Retail Management Skills Test condition to produce even slightly worse performance suggests that both conditions probably created stereotype threat. For this interpretation to be more definitive, two things would be needed: (a) the manipulation check described earlier, and (b) an effective no-stereotype threat control condition.

Ployhart et al.’s efforts to integrate research on applicant perceptions with stereotype threat research offers great potential for advancing our understanding of racial and ethnic differences in test performance. Continuing to explore how various individual differences can influence susceptibility to stereotype threat is an extremely important direction for this research program. In particular, the suggestion in this study that the degree of in-group identification among Black participants may moderate their susceptibility to stereotype threat is an important issue to pursue further.

Nguyen, O’Neal, and Ryan

Nguyen et al. had participants take a simulated personnel selection test in a 2 (race: Black, White) × 2 (threat: threat, control) between-subjects design. The test was actually a cognitive ability test composed of verbal, quantitative, and logical reasoning sections. Employing a clever cover story, the researchers convinced participants that the study was actually examining how job desirability could influence job applicants selection test performance. To help motivate the participants to take the task seriously, participants were asked to imagine that they were applying for their ideal job and that to be hired they had to “take some screening test and do better than other job applicants.” To further enhance their level of motivation, participants were informed that the top 20% of performers would receive a $10 bonus. Participants in the threat condition were then told that the personnel selection test they were about to take would provide the employer with a “genuine diagnostic evaluation of your Math, Verbal, and Logical reasoning abilities and limitations.” Although participants in the control condition were simply told they were about to take a selection test. Then immediately before taking the cognitive ability test, par-
Participants were asked to complete a demographic questionnaire in which they reported their GPA and SAT.

In light of these procedures, there is no doubt that Nguyen et al. have created a highly engaging simulated personnel selection testing environment. There is some doubt, however, as to whether they created a control condition in which African Americans would not experience stereotype threat. With all the reference to selection test performance, rewards for being a top performer on that test, and questions asking participants to indicate their GPA and SAT scores prior to taking an obviously cognitive ability test, there appears to be more than enough primes in the control condition to cause stereotype threat for Black participants.

Like the other studies in the issue, this one also found that Black participants underperformed in relation to White participants, this time even after the groups were equated on their SATs, and on all three subsections of the cognitive ability test. Such findings clearly fit the view that both the Threat and control conditions in this experiment imposed stereotype threat on Black participants. But again, this interpretation cannot be definitive without an effective no-stereotype threat control condition—an issue that should be taken up in future research.

REFERENCES


