British Journal of Social Psychology (2005), 44, 645–657
© 2005 The British Psychological Society

The British Psychological Society

www.bpsjournals.co.uk

Clearing the air: The effect of experimenter race on target's test performance and subjective experience

David M. Marx[1]* and Phillip Atiba Goff[2]

[1] Department of Social and Organizational Psychology, University of Groningen, The Netherlands

[2] Department of Psychology, The Pennsylvania State University, USA

According to stereotype threat theory (Steele, 1997), stereotyped targets underperform on challenging tests, in part because they are worried about being viewed in terms of the negative stereotype that they are intellectually inferior. How then are the negative effects of stereotype threat reduced for stereotyped targets? To examine this issue, a study was conducted to investigate whether stereotype threat's adverse effects are reduced when a Black experimenter administers a verbal test to Black participants. We further examined the question of whether Black participants have a subjective awareness of stereotype threat. Results showed that when a Black experimenter gave a verbal test to Black participants, they did not suffer the typical performance decrements associated with stereotype threat. Additionally, results supported the hypothesis that Black participants have conscious access to the experience of stereotype threat and that this effect is partially mediated by their endorsement of the stereotype.

Defining the borders and mechanisms of stereotype threat (Steele, 1997; Steele, Spencer, & Aronson, 2002) has become an increasingly important task. At its core, stereotype threat theory seeks to explain, from the target's perspective, why certain groups perform worse than their motivations and prior performances suggest they should. Particularly for women in mathematics, and Black students in most academic domains, this underperformance on important standardized tests, such as the Scholastic Assessment Test (SAT), can place them at a disadvantage in their pursuit of higher education. In the case of Black students, there are many factors that may contribute to the relative dearth of these students attending 4-year colleges. Economic hardships, inferior extracurricular opportunities, and substandard schooling at the pre-college level are just some examples of the structural barriers that many Black students must overcome to reach their academic goals. Beyond these obvious structural inequalities,

* *Correspondence should be addressed to David Marx, Department of Social and Organizational Psychology, University of Groningen, Grote Kruisstraat 2/1, 9712 TS Groningen, The Netherlands (e-mail: d.marx@ppsw.rug.nl).*

DOI:10.1348/014466604X17948

the standardized test gap is still one of the most prominent explanations for minority underrepresentation in college. Moreover, for decades this 'testing gap' has been the source of intergroup tension, political debate, and the target of countless government and local intervention programmes (Bowen & Bok, 1998; Lemann, 1999; Ramist, Lewis, & McCamley-Kenkins, 1994). Consequently, it is seen as one of the most acute problems currently facing the American education system.

Stereotype threat theory offers a unique perspective on this problem, suggesting that the traditional explanations for the testing gap (e.g. structural inequality, or more insidious explanations) may be insufficient (cf. Lemann, 1999). According to Steele (1997) there is a general 'threat in the air' whenever a negatively stereotyped group member enters a situation where negative stereotypes might apply. In evaluative situations, such as taking standardized tests, this threat can lead to underperformance for stereotyped targets, due to their concern about confirming a negative stereotype about their group (Steele, 1997; Steele & Aronson, 1995).

Despite stereotype threat theory's tremendous contributions to our understanding of the testing gap, there are at least two crucial elements of the theory that remain under-explored. The first is how one diminishes stereotype threat in high-stakes testing situations. Standard stereotype threat studies have reduced or eliminated the performance decrement, either by making the test non-diagnostic of ability (Steele & Aronson, 1995), stating that the test does not show gender differences (Spencer, Steele, & Quinn, 1999), or by changing the meaning of the test in other ways, (e.g. from a test of *athletic intelligence* to one of *athletic ability*; Stone, Lynch, Sjomeling, & Darley, 1999). While changing the social meaning of a test seems, definitionally, to be the only way to reduce stereotype threat, the problem of changing the social meaning of a diagnostic test is one to which we have found few empirical solutions (cf. Inzlicht & Ben-Zeev, 2000; Marx & Roman, 2002; Marx, Stapel, & Muller, 2005; McIntyre, Paulson, & Lord, 2002).

The second is the subjective experience of situations that induce stereotype threat. In other words, are participants aware of this ambient threat and their stereotype-related concerns or is the entire experience processed beneath awareness? There is relatively little empirical evidence from which to draw an accurate picture of participants' experiences of stereotype threat (Davies, Spencer, Quinn, & Gerhardstein, 2002; Maass & Cadinu, 2003; Schmader & Johns, 2003; Steele *et al.*, 2002; Wheeler & Petty, 2001). In this article, we address both questions because we believe that finding answers to these questions will greatly enhance our understanding of how stereotype threat affects the academic performance and subjective experience of stereotyped targets.

The experience of stereotype threat

In much of the literature, stereotype threat is defined and understood generally as the 'threatening experience' of knowing that one may be evaluated in accordance with the negative stereotypes held about one's group (Marx, Brown, & Steele, 1999; Steele, 1997; Steele *et al.*, 2002). Indeed, this situation can be triggered in a variety of ways (e.g. test diagnosticity; Steele & Aronson, 1995), but what all stereotype threat manipulations have in common is the fact that they make a stereotype about inferior intellectual ability relevant only for targets' test performance. Hence, if stereotype threat occurs because it leads to heightened accessibility of a negative stereotype, then the salience of group memberships associated with that stereotype should be greater in those settings where

the stereotype applies, in contrast to ones where it does not (e.g. Brewer & Gardner, 1996; Brewer & Weber, 1994; Steele *et al.*, 2002; Tajfel & Turner, 1979). As a result, targets in stereotype threat situations should be more sensitive to group-based information, which could then affect their subjective experience. That is, in stereotype threat situations, targets should also be more sensitive to and/or aware of group-based information that is stereotype relevant. If, for instance, the group-based information is positive (e.g. learning about or interacting with a fellow in-group member who disconfirms the negative stereotype), then it could lower targets' stereotype threat concerns (cf. Marx *et al.*, 2005).

Given this logic, we argue that the social situation itself can have profound implications for stereotyped targets, such that something as simple as who administers a test could alter the beliefs targets have about how they may be stereotyped. This may be particularly true if that test administrator is likewise perceived as competent in the stereotyped domain, because this person may provide stereotype-disconfirming information (Blanton, Crocker, & Miller, 2000; Major, Sciacchitano, & Crocker, 1993; Marx & Roman, 2002; Marx *et al.*, 2005). For example, Blanton *et al.* showed that in a stereotyped domain comparisons with competent in-group members led to assimilative effects on participants' self-esteem after receiving negative feedback, but only when the comparison target was from the same stereotyped group. Moreover, learning about a talented *in-group* member may also serve to delegitimize the stereotype; thus stereotyped targets may not endorse the stereotype about their group as much as when they are confronted with a talented *out-group* member who serves to remind them about the academic status differences between themselves and the out-group (see Blanton, Christie, & Dye, 2002; Jost & Banaji, 1994). This introduces the intriguing possibility that the presence of a Black experimenter (one who appears competent in the stereotyped domain) may reduce Black participants' endorsement of stereotypic beliefs in threatening situations. Put another way, the level of stereotype endorsement may be a direct consequence of the type of information that is presented in stereotype threat situations, such that seeing someone who provides or represents stereotype-disconfirming information may reduce targets' concerns about stereotype threat. In the end positive in-group information may affect Black participants differently than White participants, such that stereotype endorsement only serves as a mediator of Black, but not White, participants' stereotype threat scores. We examined these issues in the present research.

Hypotheses

In the current study, we tested the hypothesis that the presence of a competent Black experimenter would attenuate the effects of stereotype threat on Black participants' verbal test performance. We further hypothesized that the presence of a Black experimenter would change the participants' social reality and, thereby, the participants' experience of stereotype threat in the test-taking situation. Therefore, we made the prediction that Black participants would experience and accurately report higher levels of stereotype threat when a White experimenter, rather than a Black experimenter, administers a diagnostic test. Additionally, we predicted that stereotype endorsement would mediate the effect of the experimenter's race on Black, but not White, participants' stereotype threat scores. This is because the presence of

a Black experimenter should lessen the perceived legitimacy of the stereotype only for Black participants (Blanton *et al.*, 2002; Jost & Banaji, 1994).

Method

Participants and design

Participants were 32 Black and 27 White Harvard undergraduates who took part in exchange for pay or course credit. For this study we used a 2 (race of participant: Black, White) × 2 (race of experimenter: Black, White) between-participants design.

Participant recruitment

To assess identification with English we contacted potential participants, during e-mail screening sessions, and asked them to respond to four items about their interest and ability in English. These include (1) 'How important are writing and reading to you?' (2) 'How good are you at writing and reading?' (3) 'How important is it for you to do well on English/literature assignments?' and (4) 'How important is it for you to do well on English/literature exams?'. Participants responded by indicating a number from (1) *not at all important/very bad* to (5) *very important/very good*. Only those who responded to each of the statements with a score of three or above were eligible to participate.[1] The e-mail message further asked participants to report their verbal SAT score and the number of English/literature classes taken in college. So, in addition to assessing the participants' identification with English, we used a minimum score (610; The 83rd percentile) from the SAT, along with the requirement that each participant had taken at least one English or literature course in college. All of this was done to ensure that they had the skills to succeed on the verbal test.

Procedure

Participants reported to the laboratory individually, where they were greeted by one of four Black ($N = 29$) or one of three White ($N = 30$) experimenters who were blind to the hypotheses. Moreover, these experimenters were advanced undergraduate research assistants who helped with the study as part of a course requirement. The experimenters first made it clear that they were investigating the verbal test performance of undergraduates, and then explained to the participants that they would be taking a challenging verbal test that they had created. The experimenters also indicated that they would provide feedback about the participants' verbal ability and verbal test performance at the end of the study session (though no feedback was actually given). All of this was done to create the impression that the experimenters were verbally competent (Marx & Roman, 2002).[2] Participants then took a challenging verbal test under stereotype threat conditions. For this study we induced stereotype threat in

[1] *There were no reliable differences between the Black and White participants on the four recruitment questions.*

[2] *As a way to check on the participants' judgment of the experimenters' competence we averaged their responses to four statements, which were anchored with the terms (1) strongly disagree to (7) strongly agree. The four statements included: (1) 'The experimenter was composed', (2) 'The experimenter's score on this test would be higher than that of the average participants' score', (3) 'The experimenter was a competent test administrator', and (4) 'The experimenter's verbal SAT score would be higher than that of the average Harvard participants' score'. Results revealed no difference between the White experimenters ($M = 5.30$) and Black experimenters ($M = 5.25$), $F < 1.00$.*

two ways. First, we described the test as one that is diagnostically accurate at assessing participants' verbal strengths and weaknesses. Second, we asked all participants to indicate their race on the top of their test booklet. Each of these procedures has successfully activated stereotype threat in previous research (Steele & Aronson, 1995).

Verbal test performance

Participants were given 25 minutes to complete the verbal test. The test format resembled a typical Graduate Record Exam (GRE) verbal section and consisted of 28 problems taken directly from previous GRE exams. Problems were selected based on the percentage of participants from an earlier sample who answered those problems correctly (all questions fell within the range of 10–49%; Educational Testing Service, 1994). Verbal test performance was based on the number of problems each participant answered correctly.[3]

Study feedback form

As a way to collect the participants' thoughts about the study, as well as to administer the manipulation check, we gave them (in a sealed envelope) a study feedback form. This form was given after the participants completed the test, and contained four statements that were specifically aimed at measuring stereotype threat, and one statement about endorsement of the cultural stereotype about Blacks. In addition to these key statements, we included a number of statements to assess the participants' task motivation.[4]

Stereotype threat

The participants were asked to respond to four statements (Cronbach's $\alpha = .80$) that were intended to measure their experience of stereotype threat. Participants responded to each of the statements ('I worry that my ability to perform well on standardized tests is affected by my race'; 'I worry that if I perform poorly on this test, the experimenter will attribute my poor performance to my race'; 'I worry that people's evaluations of me will be affected by my race'; 'I worry that, because I know the racial stereotype about Blacks and scholastic achievement, my anxiety about confirming that stereotype will negatively influence how I perform on scholastic tests'), on a scale from (1) *strongly disagree* to (7) *strongly agree*. In order to assess how much stereotype threat the participants experienced, we first averaged the four stereotype threat questions from the study feedback form, and then used this score for all subsequent analyses. A high score would indicate greater levels of stereotype threat.

[3] *Immediately after the test, but before the study feedback form, participants were given a word-fragment task to assess feelings of self-doubt and stereotype activation.*

[4] *To examine participants' motivation we asked them to respond to three statements: (1) 'I was motivated to do my best on this test', (2) 'I would be interested in seeing my score on this test', and (3) 'I treated this test much like I would have a class exam or standardized test'. The scale was labelled on the ends with the terms (1) strongly disagree and (7) strongly agree. Results revealed that there were no motivational differences as function of participant race, researcher race, or the interaction of these two factors, ps* $> .26$.

Endorsement of the stereotype

To examine the participants' endorsement of the stereotype about Blacks and academic ability, they responded to the following statement: 'Though it may not be their fault, Blacks cannot perform as well as non-minorities in school.' The scale was anchored on the endpoints with the terms (1) *strongly disagree* and (7) *strongly agree*, with higher numbers indicating more endorsement.

Manipulation check

The last page of the study feedback form contained the manipulation check, which was designed to assess whether the participants correctly remembered the race of the experimenter, as this factor is particularly crucial to our hypotheses about the differential effect of experimenter race on Black participants' verbal test performance. For this purpose, participants were simply asked to indicate the experimenter's race.

Results

Excluded data and experimenter effects

The data from two Black and three White participants were excluded because they misidentified the experimenter's race. Data from one Black and three White participants were not analysed because they did not complete all the study measures.[5] We also checked for experimenter effects. Results showed that there were no differences on any of the dependent measures among the four Black experimenters, Fs < 2, as was the case among the three White experimenters, Fs < 2, thus no further mention of experimenter effects will be made.

Verbal test performance

Consistent with previous research (Steele & Aronson, 1995), we intended to measure the effect of our conditions on participants' verbal test scores while controlling for previous experience and ability. To do this, we first computed a standardized composite variable of each participant's prior number of literature/English classes taken and verbal test-taking skills (i.e. verbal SAT) and then used this composite variable as a covariate for all subsequent analyses. Next we submitted the participants' verbal test scores to a 2 (race of participant) $\times$ 2 (race of experimenter) ANCOVA (see Table 1). Results revealed an effect for the covariate, $F(1, 45) = 19.52$, $p < .01$, $\eta = .55$, a marginally reliable main effect for race of participant, $F(1, 45) = 3.58$, $p = .065$, $\eta = .27$, indicating that White participants performed better than Black participants, and a marginally reliable race of participant by race of experimenter interaction, $F(1, 45) = 3.60$, $p = .064$, $\eta = .27$. Although the expected interaction did not reach conventional levels of reliability, the pattern of cell means was consistent with our hypothesis, such that Black participants who were given the test by a White experimenter performed worse than participants in any other condition. Moreover, since our hypothesis was quite clear about the effect of the experimenter's race on Black participants' verbal test scores, we conducted simple effects tests.

[5] *We excluded participants if they did not complete all the measures because it was quite clear that they did not take the testing situation or the study seriously.*

Table 1. Mean adjusted (*SD*) verbal test performance, problems answered, verbal test accuracy, stereotype threat score, and stereotype endorsement as a function of race of experimenter and race of participant

	Race of experimenter			
	Black		White	
Race of participant	Black	White	Black	White
Verbal test performance	12.08 (2.73)	12.39 (4.33)	9.80 (3.31)	13.16 (3.01)
Problems answered	24.75 (3.52)	25.28 (2.73)	23.36 (3.56)	26.43 (1.38)
Verbal test accuracy	49% (12.62)	47% (13.54)	41% (11.78)	50% (10.93)
Stereotype threat score	2.30 (1.05)	1.76 (0.45)	3.48 (1.20)	1.49 (0.73)
Stereotype endorsement	1.40 (0.74)	1.99 (1.41)	2.22 (1.53)	1.77 (1.69)

Results show that Black participants who were given the test by a Black experimenter ($M = 12.08$, $SD = 2.73$) reliably out-scored those Black participants who were given the test by a White experimenter ($M = 9.80$, $SD = 3.31$), $F(1, 45) = 4.98$, $p < .05$, $\eta = .32$. White participants' performance results did not reliably differ between the White ($M = 13.16$, $SD = 3.01$) and Black experimenters ($M = 12.39$, $SD = 4.33$), $F(1, 45) = 0.39$, $p = .54$, $\eta = .09$. Consistent with typical stereotype threat effects, we also found that Black participants ($M = 9.80$, $SD = 3.31$) performed worse than did White participants ($M = 13.16$, $SD = 3.01$) when a White experimenter administered the test, $F(1, 45) = 10.07$, $p < .01$, $\eta = .43$. As a last test of our hypothesis we compared the verbal test scores of Black ($M = 12.08$, $SD = 2.73$) and White participants ($M = 12.39$, $SD = 4.33$) who took the test from a Black experimenter, with the results showing that there was virtually no difference between these participants' test scores, $F(1, 45) = 0.07$, $p = .79$, $\eta = .04$. In sum, the experimenter's race only made a difference for Black participants, such that when a Black experimenter administered the test Black participants performed as well as White participants, but they underperformed when given the test by a White experimenter.

Problems answered and verbal test accuracy

Having established an effect of the experimenter's race on Black participants' test scores, we turned to the question of whether this effect was caused by a decrease in the number of problems answered or to a decrease in their test accuracy (see Table 1). Accordingly, we conducted two separate ANCOVAs on the number of problems answered and verbal test accuracy (computed by dividing the number of problems answered correctly by the total number of problems answered). The first ANCOVA on the number of problems answered did not reveal any reliable effects, ps $> .15$. Consistent with previous research (Marx & Roman, 2002; Steele & Aronson, 1995), this analysis suggests that any performance decrement produced by our manipulation was not due to decreased effort or number of problems answered.

The second ANCOVA examining verbal test accuracy indicated that the race of participant by race of experimenter interaction was marginally reliable, $F(1, 45) = 2.98$, $p = .09$, $\eta = .25$ (other Fs < 1). Simple effects tests within race of participant by race of experimenter revealed, as expected, that Black participants were more accurate when they were given the test by a Black experimenter ($M = 49\%$, $SD = 12.62$) compared to

a White experimenter ($M = 41\%$, $SD = 11.78$), $F(1, 45) = 4.64$, $p < .04$, $\eta = .31$. White participants' verbal test accuracy was not reliably affected by the experimenter's race ($M_{\text{White}} = 50\%$, $SD = 10.93$; $M_{\text{Black}} = 47\%$, $SD = 13.54$), $F(1, 45) = 0.22$, $p = .64$, $\eta = .07$. In addition, when we compared the verbal test accuracy of Black participants ($M = 41\%$, $SD = 11.78$) relative to White participants ($M = 50\%$, $SD = 10.93$) who were given the test by a White experimenter we found a typical stereotype threat effect, $F(1, 45) = 4.86$, $p = .03$, $\eta = .31$. Yet, no differences emerged between the White ($M = 47\%$, $SD = 13.54$) and Black participants ($M = 49\%$, $SD = 12.62$) when a Black experimenter administered the test, $F(1, 45) = 0.19$, $p = .67$, $\eta = .06$. Taken together these results suggest that it is decreased accuracy, and not decreased effort, as would be shown by a difference in the number of problems answered, that contributed to Black participants having lower verbal test scores when a White experimenter administered the test.

Stereotype threat

Recall our hypothesis that Black participants would experience, and importantly report, increased feelings of stereotype threat under heightened levels of stereotype awareness. For the purposes of the present research, this means that Black participants should report higher levels of stereotype threat when a White experimenter, rather than a Black experimenter, administered the test (see Table 1). The participants' stereotype threat scores were analysed using a 2 (race of participant) × 2 (race of experimenter) ANCOVA. Results revealed main effects for the covariate, $F(1, 45) = 4.17$, $p < .05$, $\eta = .29$, and race of participant, $F(1, 45) = 15.10$, $p < .01$, $\eta = .50$, such that Black participants felt more stereotype threat than did White participants. As expected, the interaction was also reliable, $F(1, 45) = 7.21$, $p = .01$, $\eta = .37$.

Next we performed simple effects tests. According to our hypothesis, White participants should not experience reliably different levels of stereotype threat regardless of whether a White ($M = 1.49$, $SD = 0.73$) or a Black experimenter ($M = 1.76$, $SD = 0.45$) gave them the test, and that is what we found, $F(1, 45) = 0.42$, $p = .52$, $\eta = .10$. For Black participants, however, the experimenter's race should make a difference. Results from this analysis revealed strong support for this hypothesis showing that Black participants experienced, and more important, reported, higher levels of stereotype threat when a White ($M = 3.48$, $SD = 1.20$) rather than a Black experimenter ($M = 2.30$, $SD = 1.05$) gave them the verbal test, $F(1, 45) = 11.78$, $p < .01$, $\eta = .46$. Follow up comparisons demonstrated that when a White experimenter administered the test, Black participants ($M = 3.48$, $SD = 1.20$) had higher stereotype threat scores compared with White participants, ($M = 1.49$, $SD = 0.73$), $F(1, 45) = 31.04$, $p < .01$, $\eta = .64$. When a Black experimenter gave the test, there was no difference between the Black ($M = 2.30$, $SD = 1.05$) and White participants' ($M = 1.76$, $SD = 0.45$) stereotype threat scores, $F(1, 45) = 1.77$, $p = .19$, $\eta = .19$. These results clearly show that the presence of a verbally competent in-group member reduced Black participants' stereotype threat scores, but this did not occur for the White participants.

In addition to the stereotype threat scale, we gave participants a word-fragment completion task similar to the one used by Steele and Aronson (1995) to assess feelings of self-doubt and stereotype activation. This task was administered immediately after the test, but before completing our study feedback form. No reliable results occurred for participants' feelings of self-doubt or stereotype activation, except for a theoretically

uninteresting interaction between race of participants and race of experimenter on participants' self-doubt scores, $F(1, 45) = 4.28$, $p < .05$, $\eta = .29$, demonstrating that White participants indicated feeling more self-doubt when the test was administered by a Black experimenter and Black participants felt more self-doubt when a White experimenter administered the verbal test (other Fs < 1). The null effects of stereotype activation may have occurred for a couple of reasons. First, the placement of the word-fragment task deviated from Steele and Aronson's research, since we administered the task after, rather than before, the test was completed. Thus, it seems likely that the stereotype about Blacks was initially activated, but by the time we measured stereotype activation the stereotype threat had dissipated (see Kunda, Davies, Adams, & Spencer, 2002, for a related argument). And second, of the 50 participants, only 29 completed the word-fragment with at least one stereotype-related word, hence we may have failed to find any meaningful effects due to low power and/or floor effects on this measure. Given these results the question still remains as to what other factors could account for this reduction in stereotype threat among Black participants.

Mediation analyses

Based on our theoretical framework and the predicted findings on our stereotype threat measure, we began to explore whether participants' level of stereotype endorsement mediated the effect of the experimenter's race on their stereotype threat scores (e.g. Blanton *et al.*, 2002). To do this we used procedures recommended by Baron and Kenny (1986).

According to this approach, three relationships between our factors must be established in order to test for mediation. First, the experimenter's race (IV) must predict the participants' stereotype threat scores (DV); second, the experimenter's race must predict their stereotype endorsement scores (the mediator), and third, their stereotype endorsement scores must predict their stereotype threat scores. If these criteria are met, then the participants' stereotype threat scores can be regressed onto the race of the experimenter and their stereotype endorsement scores in a final regression analysis. Support for mediation is obtained by demonstrating the effect of the experimenter's race on the participants' stereotype threat scores is no longer reliable when accounting for their endorsement of the cultural stereotype. We conducted the regression models separately for Black and White participants, since we expected that endorsement of the stereotype about Blacks and their inferior academic ability would only mediate the effect of the experimenter's race on Black participants' experience of stereotype threat.

For Black participants, the experimenter's race increased their endorsement of the stereotype, $\beta = 0.45$, $t(26) = 2.02$, $p = .05$, and their stereotype threat scores, $\beta = 0.58$, $t(26) = 2.83$, $p < .01$, meaning that when a White experimenter administered the verbal test, Black participants had higher stereotype endorsement scores as well as higher stereotype threat scores. The third regression analysis showed that stereotype endorsement was a reliable predictor of the Black participants' stereotype threat scores, $\beta = 0.51$, $t(26) = 3.15$, $p < .01$, such that higher endorsement scores were associated with higher stereotype threat scores. The final regression analysis showed that when the Black participants' stereotype threat scores were regressed onto the race of the experimenter and their endorsement scores, the experimenter's race was no longer a reliable predictor of the Black participants' stereotype threat scores, $\beta = 0.33$, $t(26) = 1.99$, $p = .06$, while their stereotype endorsement scores remained reliable, $\beta = 0.39$, $t(26) = 2.36$, $p < .03$. A Sobel (1982)

test of the reduction in the direct effect of the experimenter's race on participants' stereotype threat scores was marginally reliable, $Z = 1.54$, $p = .10$. Although this last effect was only marginally reliable it still indicates that for Black participants, their stereotype threat scores were caused, in part, by their endorsement of the stereotype about their group, and that the presence of a Black experimenter in a stereotype threat situation reduces the associated concern.[6]

For White participants, we found that the experimenter's race was not a reliable predictor of their stereotype threat scores, $\beta = -0.01$, $t(26) = -0.50$, $p = .62$, demonstrating that White participants' experience of stereotype threat was not sensitive to the race of the experimenter. Taken together these regression analyses support our hypothesis that endorsement of the stereotype contributes to effect of the experimenter's race on Black participant's experience with stereotype threat. Moreover, this effect provides additional evidence for one of the main tenets of stereotype threat theory: stereotype threat is unique to those individuals who are aware of and concerned about confirming the negative stereotype about their group (Marx *et al.*, 1999; Steele, 1997; Steele *et al.*, 2002).

Discussion

This study demonstrates that having a Black experimenter administer a challenging verbal test to highly motivated Black participants under stereotype threat conditions protected those participants' verbal test performance. In fact, having a Black experimenter allowed Black participants to out-score those Black participants who were given the test by a White experimenter, hence simulating the typical performance decrement associated with stereotype threat. Importantly, and consistent with previous research (Marx & Roman, 2002; Steele & Aronson, 1995), this was not a result of the Black participants answering fewer problems. Rather, it was due to a decrease in their verbal test accuracy. Moreover, Black participants appeared to be aware that a vague 'threat in the air' hovered over the test-taking situation that produced their depressed scores. This finding demonstrates that stereotype threat can be consciously accessible and that it may be caused, in some degree, by the concern Black participants have regarding the negative stereotype about their group's academic ability.

In sum, our results imply that the presence of Black experimenters altered the test-taking situation for Black participants. Specifically, a Black experimenter changed the meaning of the test in such a way that Black participants neither underperformed relative to Whites, nor felt as much of a racial threat from the test-taking situation. This result provides the first evidence that stereotype threat is a consciously accessible phenomenon, which can be measured through self-report. Moreover, we found that the effect of the experimenter's race on Black participant's stereotype threat scores was somewhat reduced by Black participants' endorsement of the stereotypes about their group. In other words, having a Black, rather than a White, experimenter administer a test may serve to delegitimize the threat of being stereotyped for Black participants.

[6] *In light of these findings one might ask whether stereotype threat mediates the relation between experimenter race and participants' verbal test performance. Unfortunately, the order in which these measures were administered precludes us from making any strong causal claims; however, we did examine within-cell partial correlations (controlling for the covariate) among these measures. Results only revealed a marginally reliable positive correlation ($r = .50$, $p = .08$) between stereotype threat and participants' verbal test performance for Black participants who were given the test by a White experimenter (other ps $> .68$), but this effect should be interpreted with caution because of low power.*

In the end the presence of stereotype-disconfirming in-group members may serve as a buffer against the proverbial hammer of negative stereotypes (Allport, 1954).

Loose ends and implications beyond the laboratory

Though the current results are encouraging, a few issues should be discussed. One primary issue is the fact that we did not include a non-diagnostic condition. Nevertheless, we believe that this design limitation does not detract from the impact of our results because Black participants generally underperform under stereotype threat conditions; thus this condition was central to our hypotheses about the benefits of a Black experimenter as well as to real-world effects (Bowen & Bok, 1998). Also, because we had a limited Black participant population we felt that the best use of the participants we did have would be to use them in what we viewed as the most critical situation. However, because there was no true 'non-diagnostic' condition, several hypotheses cannot be ruled out by the present research. For instance, we are not able to conclude how the presence of a competent Black experimenter affects Black participants' verbal test performance when the negative stereotype is not activated in the testing situation. Although we have argued that the presence of a verbally competent Black experimenter reduces ambient threat, it may be that a competent in-group member provides a positive social comparison for Black participants who are in a stereotype threat situation (Blanton *et al.*, 2000; Schmitt, Silvia, & Branscombe, 2000), which then could lead to improved performance even in non-threatening situations (e.g. Dijksterhuis *et al.*, 1998). As compelling as this explanation might be, it does seem that our results are more than just assimilation effects. Because we found reliable differences on our threat measure, it seems that the presence of a competent Black experimenter makes the cultural stereotype less threatening. Similarly, since the presence of a Black experimenter decreased stereotype endorsement for Black participants, which in turn reduced their stereotype threat scores, it seems that our manipulations were clearly relevant to participants' experience of threat, rather than simply due to positive in-group comparisons.

One other point that deserves note is whether a stereotyped group member must be in charge, or whether simply learning about a minority group member who disconfirms the negative stereotype still serves the same purpose. Although the present research does not speak to this point, other research may help to clarify this issue. For example, in a study by Marx, Urland, Overbeck, and Webster (2002) it was shown that when female participants learn about a mathematically-talented woman from the same university who is applying for a mathematics tutor position they perform better, even under evaluative conditions, compared with when the job candidate is not so highly talented in mathematics. Importantly this occurred even though the job candidate was fictitious and not part of the immediate testing situation. So, in a sense, having a Black experimenter in charge of the study may be a sufficient, but not a necessary, condition to alleviate the standard stereotype threat performance decrement.

Clearly additional research is needed to determine the boundary conditions of our effects. However, this study provides a possible strategy for researchers and educators to change the social reality of high-stakes evaluative situations. Importantly, this intervention did not depress the scores of White participants (see Danso & Esses, 2001, for related research showing that White participants perform better when a Black experimenter administers a test), suggesting that the presence of a competent group member, such as a teacher, may reduce stereotype threat for stereotyped targets, while leaving the performance of non-stereotyped targets untouched. Taken together, we

suggest that eliminating stereotype threat in diagnostic conditions may be a less daunting task than previously feared, and this makes for good theoretical and practical news.

Acknowledgements

The research reported in this article formed part of the first author's doctoral dissertation. This study was conducted at Harvard University, supported by grants from the Stimson Fund (33-440-2575-2) and the Knox Fund (33-440-18102-30) at Harvard University, and a Grants-in-Aid award from the Society for the Psychological Study of Social Issues (SPSSI) to the first author. Portions of this article were written while the first author was a postdoctoral fellow at the University of Colorado, Boulder. We thank Andrew Amo, Terry-Ann Burrell, Anwar Floyd-Pruitt, Mateo Jaramillo, Chris Jenkins, Chris Soto, Josh Weaver, and Megan Whyte for their able assistance with the execution of this study. We also thank the laboratory of Claude M. Steele and Benoit Monin, as well as Paul G. Davies, Ernestine Gordijn, Hana Rae Shepherd, and Vincent Yzerbyt for their helpful comments on earlier versions of this article. Finally, we want to thank Terri Vescio for her invaluable advice regarding this article.

References

Allport, G. W. (1954). *The nature of prejudice*. New York: Doubleday Books.

Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology, 51*, 1173–1182.

Blanton, H., Christie, C., & Dye, M. (2002). Social identity versus reference frame comparisons: The moderating role of stereotype endorsement. *Journal of Experimental Social Psychology, 38*, 253–267.

Blanton, H., Crocker, J., & Miller, D. T. (2000). The effects of in-group versus out-group social comparison on self-esteem in the context of a negative stereotype. *Journal of Experimental Social Psychology, 36*, 519–530.

Bowen, W. G., & Bok, D. C. (1998). *The shape of the river: Long-term consequences of considering race in college and university admissions*. Princeton, NJ: Princeton University Press.

Brewer, M. B., & Gardner, W. (1996). Who is this "we"? Levels of collective identity and self-representations. *Journal of Personality and Social Psychology, 71*, 83–93.

Brewer, M. B., & Weber, J. G. (1994). Self-evaluation effects of interpersonal versus intergroup social comparison. *Journal of Personality and Social Psychology, 66*, 268–275.

Danso, H. A., & Esses, V. M. (2001). Black experimenters and the intellectual test performance of White participants: The tables are turned. *Journal of Experimental Social Psychology, 37*, 158–165.

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.

Dijksterhuis, A., Spears, R., Postmes, T., Stapel, D., Koomen, W., Knippenberg, A., & Scheepers, D. (1998). Seeing one thing and doing another: Contrast effects in automatic behavior. *Journal of Personality and Social Psychology, 75*, 862–871.

Educational Testing Service (1994). *GRE: Practicing to take the general test* (9th ed.). Princeton, NJ: Author.

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.

Jost, J. T., & Banaji, M. R. (1994). The role of stereotyping in system-justification and the production of false consciousness. *British Journal of Social Psychology, 33*, 1–27.

Kunda, Z., Davies, P. G., Adams, B. D., & Spencer, S. J. (2002). The dynamic time course of stereotype activation: Activation, dissipation, and resurrection. *Journal of Personality and Social Psychology, 82*, 283–299.

Lemann, N. (1999). *The big test: The secret history of the American meritocracy*. New York: Farrar, Straus and Giroux.

Maass, A., & Cadinu, M. (2003). Stereotype threat: When minority members underperform. In W. Stroebe & M. Hewstone (Eds.), *European review of social psychology* (Vol. 14, pp. 243–275). New York: Taylor and Francis Group.

Major, B., Sciacchitano, A. M., & Crocker, J. (1993). In-group vs. out-group comparisons and self-esteem. *Personality and Social Psychology Bulletin*, *19*, 711–721.

Marx, D. M., Brown, J. L., & Steele, C. M. (1999). Allport's legacy and the situational press of stereotypes. *Journal of Social Issues (Prejudice and Intergroup Relations: Papers in Honor of Gordon W. Allport's Centennial)*, *55*(3), 491–502.

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

Marx, D. M., Stapel, D. A., & Muller, D. (2005). We can do it: The interplay of construal orientation and social comparisons under threat. *Journal of Personality and Social Psychology*, *88*, 432–446.

Marx, D. M., Urland, G. R., Overbeck., J. R., & Webster, G. D. (2002). *Upward social comparisons in a stereotyped domain: The effect of female role models on students' feelings of doubt and math performance*. Unpublished manuscript, University of Colorado, USA.

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

Ramist, L., Lewis, C., & McCamley-Kenkins, L. (1994). Implications of using freshman GPA as the criterion for the predictive validity of the SAT. In W. W. Willingham & C. Lewis (Eds.), *Predicting college grades: An analysis of institutional trends over two decades* (pp. 253–288). Princeton, NJ: Educational Testing Service.

Schmader, T., & Johns, M. (2003). Converging evidence that stereotype threat reduces working memory capacity. *Journal of Personality and Social Psychology*, *85*, 440–452.

Schmitt, M. T., Silvia, P. J., & Branscombe, N. R. (2000). The intersection of self-evaluation maintenance and social identity theories: Intragroup judgment in interpersonal and intergroup contexts. *Personality and Social Psychology Bulletin*, *26*, 1598–1606.

Sobel, M. E. (1982). Asymptotic confidence intervals for indirect effects in structural equation models. In S. Leinhardt (Ed.), *Sociological Methodology* (pp. 290–312). Washington, DC: American Sociological Association.

Spencer, S. J., 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., & Aronson, J. (1995). Stereotype vulnerability 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* (Vol. 34, pp. 379–440). San Diego, CA: Academic Press.

Stone, J., Lynch, C. I., Sjomeling, M., & Darley, J. M. (1999). Stereotype threat effects on Black and White athletic performance. *Journal of Personality and Social Psychology*, 77, 1213–1227.

Tajfel, H., & Turner, J. C. (1979). An integrative theory of intergroup conflict. In W. Austin & S. Worchel (Eds.), *The social psychology of intergroup relations* (pp. 33–48). Pacific Grove, CA: Brooks.

Wheeler, S. C., & Petty, R. E. (2001). The effects of stereotype activation on behavior: A review of possible mechanisms. *Psychological Bulletin*, *127*, 797–826.

Received 15 January 2003; revised version received 10 October 2004