

RUNNING HEAD: RACIAL PARALYSIS

Racial Neutrality and Racial Paralysis

Michael I. Norton

Harvard Business School

Malia F. Mason

Columbia Business School

Joseph A. Vandello & Andrew Biga

University of South Florida

Rebecca Dyer

Yale University

Abstract

Four studies examine the existence, underlying mechanism, and effectiveness of a new norm endorsed by both Black and White Americans for managing interracial interactions: “racial paralysis,” the tendency to opt out of decisions involving members of different races. While Whites were quite willing to choose which of two White individuals was more likely to be class valedictorian or to have committed a violent crime, they were less likely to make the same choice between a White and Black person (Study 1). Study 2 examined the strength of this tendency to opt out; Whites were willing to forgo a monetary incentive to avoid choosing. Study 3 used fMRI to examine the mechanisms underlying racial paralysis, revealing greater recruitment of brain regions implicated in conflict in social decision-making, and inhibition of instinctively preferred but contextually inappropriate responses when making cross-race choices. Finally, Study 4 explored the effectiveness of this strategy, demonstrating that both White and Black Americans view opting out as an effective means of appearing unbiased. We discuss the impact of racial paralysis on the quality of interracial relations.

.

Racial Neutrality and Racial Paralysis

Imagine stepping onto a crowded subway car, shopping bags in each hand, and finding two seats left, each next to a similarly dressed man: one White, the other Black. Where would you sit? If you are White, choosing to sit next to the White passenger raises the concern that you will be seen as biased, while choosing to sit next to the Black passenger raises the concern that you will be seen as – perhaps disingenuously – bowing to political correctness. Nor does being Black solve the dilemma; even for a Black passenger, either decision appears to constitute a choice made on the basis of race. What happens in these common situations, when individuals must decide in a split second who to sit next to on a bus, who to ask for directions, or who to stand next to in an elevator? Even more problematically, what happens when such situations come with increased consequences, such as in discussions about who to hire or admit to college: a White or a Black candidate?

We suggest that the concern about appearing biased elicited by such situations creates conflict about the appropriate response. As a result, one popular – if sometimes suboptimal – solution is to opt out of the decision altogether in an effort to display racial neutrality: Despite the weight of their shopping bags, individuals may choose to forgo either seat and remain standing, rather than risk the appearance of bias. We suggest that similar solutions to such problems are representative of an emerging trend in interracial relations, which we term racial paralysis: The tendency for people to opt out of situations that require choices seemingly made on the basis of race. While such situations can come with immediate costs to the individual – coping with the stress of making such decisions, or remaining standing in our example – they can also have broader and more-long term costs, in the form of decreased interracial interaction.

Trends in Interracial Relations

Myrdal (1944) identified relations between Whites and Black Americans as the “problem of the century,” and indeed social scientists have been documenting trends in anti-black prejudice for nearly a century. A long tradition of research has explored Americans’ evolving perceptions of anti-Black bias (e.g., Crosby, Bromley, & Saxe, 1980; Cuddy, Fiske, & Glick, 2008; Devine & Elliot, 1995; Katz & Braly, 1933; McConahay, 1986; Nosek, Banaji, & Greenwald, 2002); in general, these surveys have documented a slow decline in expressions of overt racism against Blacks. At the same time, however, several streams of research have demonstrated that racism still exists, albeit in more subtle forms.

Most recently, a growing body of research has demonstrated that while explicit attitudes towards Blacks have become more positive over time, implicit measures of those attitudes continue to reveal bias (Devine, 1989; Fazio, Jackson, Dunton, & Williams, 1995; Greenwald & Banaji, 1995; Greenwald, McGhee, & Schwarz, 1998), which correlate with biased behavior (e.g., Correll, Park, Judd, & Wittenbrink, 2002; Dovidio, Kawakami, & Gaertner, 2002; Jost et al., 2009). The experience of the participant in a standard Implicit Association Task study is informative for understanding how interracial situations induce a feeling of racial paralysis: During the test, participants frequently notice the difficulty they are having pairing positive words with Black faces, especially when compared to the ease with which they pair those same words with White faces. On receiving their scores (which frequently indicate pro-White implicit bias) participants are faced with the fact that while they thought they were egalitarian, they in fact may harbor biased racial attitudes. Importantly for our account, while many participants would find demonstrating a pro-Black bias more comforting than a pro-White bias, most participants would be happiest if their scores showed them to be race-neutral; indeed, when

participants attempt to “fix” their performance, they often do so by trying to equalize their reaction times for the different versions of the task.

When making decisions between members of different races, people also seek to appear race-neutral; in research on aversive racism, for example, while Whites continue to exhibit bias against Blacks, they do so only when able to justify that behavior to themselves and others (Dovidio & Gaertner, 2004; Gaertner & Dovidio, 1986; see also Crandall & Eshleman, 2003; Snyder, Kleck, Strenta, & Mentzer, 1979). Even in the rare cases in which people show favoritism towards Blacks, decision-makers are still likely to claim that race was not a factor. In Hodson, Dovidio, and Gaertner (2002), for example, Whites who favored White candidates for admission to college over similarly-qualified Black candidates chose other criteria to justify their decisions (e.g., their preferred candidate’s GPA); Whites who favored *Black* candidates over similarly-qualified *White* candidates, ironically, engaged in the same strategy, using non-racial criteria to justify their choice of the Black candidate to the same extent as those who favored the White candidates (see also Norton, Vandello, & Darley, 2004; Saucier, Miller, & Doucet, 2005).

More generally, we suggest that despite their seeming differences, the large body of research demonstrating that people will behave *negatively* towards Blacks only when they have some available justification and the few investigations that demonstrate *favoritism* towards Blacks have a common underlying theme: discomfort in making decisions that involve members of different races. Whether allowing a poorly dressed Black patron to enter a restaurant for fear of appearing biased, or refusing to help a Black person when one can justify it (Dovidio & Gaertner, 1981), interracial situations increasingly evoke feelings of uncertainty stemming from a desire to appear unbiased. Indeed, several measures of individual differences in prejudice have as a core component people’s desire to appear unbiased – both to others and to themselves (Dunton & Fazio, 1997; Plant & Devine, 1998). We suggest that in current American culture,

people can be less concerned with expressing their bias *against* Blacks or demonstrating their lack of bias *toward* Blacks than merely wishing to appear as though they have no preference at all.

This desire for racial neutrality has become increasingly prevalent in American culture, as reflected by the recent attention given by sociologists and psychologists to the emergence and ramifications of colorblindness as a means of dealing with interracial relations (e.g., Bonilla-Silva, 2003; Pager & Quillian, 2005; Plaut, Thomas, & Goren, 2009; Richeson & Nussbaum, 2004; Wolsko, Park, Judd, & Wittenbrink, 2000). In one investigation, for example, Whites were asked to complete a task that required describing photographs to another person; Whites who played with a Black partner frequently avoided mentioning race, even when race was a highly diagnostic feature, an omission that negatively impacted their performance on the task (Norton, Sommers, Apfelbaum, Pura, & Ariely, 2006). In addition, this tendency toward colorblindness appears to be socially constructed; when alone, people were quite facile at categorizing faces based on race, while the tendency to avoid race was exacerbated when norms of colorblindness were made salient by other players (Apfelbaum, Sommers, & Norton, 2008).

This trend toward racial neutrality is also evidenced in two domains that directly impact public life: legal discourse and educational philosophy. While Regents of the University of California v. *Bakke* (1978) affirmed the right of universities to give preference to minority applicants to foster diversity, the Supreme Court has moved increasingly toward a view that both racism against minorities and affirmative action in favor of minorities are biased (Carbado & Harris, 2008); in *Ricci v. DeStefano* (2009), for example, the Court ruled that New Haven, CT had discriminated against *White* firefighters by favoring Black firefighters for promotion. In the domain of elementary and secondary education, as well, a desire for racial neutrality has taken root in recent decades. Pollock (2004), for example, notes the ubiquity of phrases such as “race

does not matter” and “we are all the same” in teaching and talking about race (see also Schofield, 2007). Indeed, evidence suggests that while young children do not instinctively adopt a colorblind practice, they do internalize these norms over time. In one investigation, while young children (ages 8-9) were very willing to use racial descriptors when describing others, this tendency decreased dramatically by ages 10-11 – precisely the moment when children become sensitive to cultural norms (Apfelbaum, Pauker, Ambady, Sommers, & Norton, 2008).

Racial Neutrality and Racial Paralysis

Of course, teaching children that race should not be used in judging others derives from a noble impulse to impart values of fairness and equity. What are the consequences, however, of the emergence and endorsement of racial neutrality – of not noticing or mentioning race – for the nature and quality of interracial relations? While norms of colorblindness likely arose from well-meaning intentions – “the best way to be egalitarian is to not even notice race” – the norms provide very little guidance in everyday situations, such as the subway situation with which we opened. If showing any preference in any situation can be construed as evidence of bias, how should a person in a diverse setting behave? We suggest that norms of racial neutrality can in some situations induce “racial paralysis,” where people’s concern with appearing unbiased can inhibit both what they say and what they do – all in the direction of saying and doing nothing, but rather opting out of such situations altogether.

We use a paradigm that captures the most basic form of this dilemma: Forgoing a choice between two individuals of different races solely on the basis of photographs of their faces. Such an unwillingness to judge faces would stand in stark contrast to people’s skill at face perception (e.g., Chernoff, 1973; Smith, Cottrell, Gosselin, & Schyns, 2005; Valenza, Simion, Cassia, & Umiltà, 1996; Zebrowitz, 1997) and willingness to make judgments on that basis. As just one

example, people are quick to form judgments about facial attractiveness (Willis & Todorov, 2006), and associate a host of positive traits with attractive individuals (Nisbett & Wilson, 1977) which then guides more positive behavior toward them (Snyder, Tanke, & Berscheid, 1977; Solnick & Schweitzer, 1999; Zebrowitz & McDonald, 1991). Even more relevant to the present investigation, people are also willing to choose *between* individuals on the basis of attractiveness (Johansson, Hall, Sikstrom, & Olsson, 2005); indeed, people generally are comfortable making choices between people based on their faces on a variety of dimensions, such as which of two individuals is more likely to be a member of a given profession (Hassin & Trope, 2000).

In the example with which we opened, however, all of these fine-tuned processes appear to come to a crashing halt: In particular, we suggest that choosing between two individuals from different racial groups – which in theory employs many of the same processes as choosing between members of the same groups – is in practice something that people are loathe to do. It is not that judging people based on their race is inherently more difficult, since categorizing people by their race is a relatively effortless task (Ito & Urland, 2003; Montepare & Opeyo, 2002), and people do draw inferences about members of other racial groups based on their photographs (e.g., Blair, Judd, Sadler, & Jenkins, 2002). Instead, we suggest that while choosing between two faces of the same race constitutes mere perceptual discrimination between those individuals, choosing between members of different races has greater significance – due to the concern that any decision may serve as evidence of bias – and therefore induces greater decision conflict, leading individuals to opt out.

Overview of the Studies

We first demonstrate White participants' willingness to make choices between individuals of the same race and the racial paralysis they experience when asked to make cross-

race judgments, in both negative and positive domains (Study 1). We next assess the magnitude of this reluctance to choose, pitting the desire to opt out against monetary incentives for choosing (Study 2). In Study 3, we both document conditions under which racial paralysis is most likely to occur, and examine the psychological mechanisms underlying decisions to opt out using functional magnetic resonance imaging (fMRI). Finally, Study 4 assesses the efficacy of racial paralysis by examining whether a national sample of both Black and White Americans agree that opting out accomplishes the decision makers' goal of appearing unbiased.

Study 1: Pick the Criminal, Pick the Valedictorian

We first wanted to establish that people are less willing to choose between members of different races than members of the same race. Importantly, we predict a different pattern of results than either research suggesting that people are biased against Blacks (which might suggest that people would make more positive judgments about Whites) and research suggesting that people can favor Blacks in a desire to appear unbiased (which might suggest that people would make more positive judgments about Blacks). We predict that emerging norms of racial neutrality make picking *either* a White person or a Black person inappropriate, leading people not to favor members of one race over another, but instead to opt out of decisions altogether. In addition, while it is easy to imagine that people might be unwilling to express judgments in negatively-valenced domains (e.g., which person is more violent), our account – and our opening example – suggests that unwillingness to choose should occur in both positive and negative domains, since making choices between members of different races violates norms of racial neutrality regardless of the specific judgment.

Method

Participants ($N = 107$) were White undergraduates who were approached in campus centers and dining halls and asked to take part in a brief survey testing “gut feelings.” Participants were told that gut feelings were “automatic, emotional responses to stimuli” and that there were no right or wrong answers.

In the *Criminal* task, they were then shown two faces and indicated which person they thought had committed a violent assault by checking a box underneath one of the faces or a third box labeled “I have no gut feeling.” Participants were randomly assigned to make same-race (a choice between two White males) or cross-race (a choice between a White male and a Black male) choices. See Figure 1 for sample stimuli.

In the *Valedictorian* task, participants were asked to choose which of two people they thought would perform better in college. In this version, they again saw either two White faces (same-race) or one White and one Black face (cross-race); we added an additional same-race condition with two Black faces, to address two alternative explanations. First, it is possible that Whites might fail to make a choice in the cross-race condition not due to the different races but rather because the presence of any Black face in the array makes choice suspect; our account, however, holds that refusal to choose to occur only when faces are of different races. Second, it is also possible that a failure to choose between a White and Black face is due to Whites’ relative lack of familiarity with Black faces (Malpass & Kravitz, 1969), making judgments about such faces more difficult; if this were the case, then judgments between two Black faces would be particularly difficult, while our account suggests that these judgments are relatively easy.

Results and Discussion

Participants who saw two Whites in the *Criminal* task were willing to choose one of the two faces, as 80% did so. When one face was White and the other Black, however, just 52%

chose, $\chi^2(1) = 4.17, p < .05$; thus while just one-fifth of participants claimed to have no gut feeling when making same-race choices, nearly half did so when making cross-race choices.

Similarly, participants were quite willing to choose a face in the *Valedictorian* task when both faces were White (84%), and were equally willing when both were Black (85%), suggesting that a failure to choose in the cross-race condition is not due to an overall tendency for Whites to forgo choice in the presence of minority faces or an inability to differentiate between members of minority groups. However, as predicted, fewer participants (55%) were willing to choose in the cross-race judgment, $\chi^2(2) = 6.12, p < .05$.

We also examined the choice shares for the Black and White candidate among those participants who did choose in the cross-race decisions. More participants chose the Black face (45%) than the White face (10%) in the positively-valenced *Valedictorian* task, seemingly suggesting a bias in favor of Blacks; however, the Black face (35%) was also chosen more than the White face (17%) in the negatively-valenced *Criminal* task as well. Most important for our account, results from both tasks suggest that the dominant response when choosing between faces of different races is to opt out altogether.

While forgoing choice serves as an indicator that choice is aversive (Dhar & Simonson, 2003; Larrick, 1993), we sought additional evidence of this discomfort by examining whether people would go to greater lengths to justify such choices. We asked an additional sample of White undergraduates ($N = 100$) to complete the *Valedictorian* task and required them to provide written explanations for their choices. Not only were participants again more likely to choose between two White faces (84%) than one White and one Black face (60%), $\chi^2(1) = 7.18, p < .01$, they went to greater lengths to explain their decisions – by using more words to explain them – when choosing between one White and one Black face ($M = 19.20, SD = 18.71$) than two White faces ($M = 12.36, SD = 10.21$), $t(98) = 2.20, p < .04$. In the same-race condition, for example,

one participant actually used stereotypes – albeit more acceptable stereotypes about a White ethnic group – to justify his choice, writing: “The guy on the right looks somewhat Irish → drinker → party boy → will do worse in school.” In contrast, a participant in the cross-race condition ended his lengthy explanation with: “Totally impossible to judge people based on pictures!”

When participants were asked to give their gut reactions about personal characteristics (both positive and negative) based solely on people’s faces, the overwhelming majority were willing to do so – unless the two faces pictured were of different races. Participants were more likely to opt out of such cross-race decisions, demonstrating racial paralysis.

Study 2: The Costs of Racial Paralysis

Study 1 and the follow-up study suggested that Whites are uncomfortable choosing between members of different racial groups, but how strong is this desire to appear unbiased? Previous research has demonstrated that people will choose between members of different groups when such choices can be masked with a non-racial explanation (Hodson et al., 2002; Norton et al., 2004). To return to our opening example, if one’s shopping bags were heavy enough, one might be seen as justified in sitting next to anyone rather than bear the cost of standing overburdened. Given that monetary incentives have been shown to decrease the impact of other social norms such as conformity (Baron, Vandello, & Brunsman, 1996), we chose to use different levels of monetary incentives to examine the level of incentive required for people to forgo racial neutrality and make a choice. In short, we benchmark the strength of the desire for racial neutrality against the cost of forgoing cash.

Method

Participants ($N = 60$) were White undergraduates approached in a campus student center. While participants made only one judgment in the previous studies, in this study we added three non-diagnostic same-race choices, not only to mask the purpose of the study, but also to compare the effects of incentives on these filler choices to cross-race choices. Participants indicated which of two White females they thought was a member of the marching band, which of two White males they thought was Canadian, which of two males – one White and one Black, our cross-race judgment – they thought had a perfect GPA over the last two years, and which of two White females they thought had spent a semester abroad in Italy.¹ As in Study 1, they could choose either face or indicate that they had no gut feeling.

Unlike Study 1, in which participants were explicitly told there was no right or wrong answer, in this study we told participants that their task was to be accurate in their selections. We told them we had asked the individuals in the photographs to complete surveys about themselves such that we knew, for example, which person was Canadian, and that the goal of our study was to see if people could guess information about them on the basis of their pictures. To motivate their performance, we randomly assigned participants to one of three incentive levels: no incentive (as in Study 1), \$1 for each correct answer, or \$5 for each correct answer. A failure to choose a candidate in either of the two incentives conditions, therefore, meant forgoing a fifty percent chance at \$1 or \$5.

Results and Discussion

For the three filler choices between two White faces (marching band, Canadian citizenship, semester in Italy), incentives had little impact. Because participants' likelihood of choosing a candidate was very high even in the absence of incentives, there was no room for incentives to increase choice, all $\chi^2 < 2.15$, $ps > .34$: the percentage of participants choosing a candidate across the three same-race tasks with no incentives was 93%, which increased to 95%

with the \$1 incentive, and 98% with the \$5 incentive. When participants chose between a White and a Black candidate in the GPA judgment, however, incentives had a large impact. As before, participants were relatively unwilling to make cross-race choices with no incentive (just 65%, compared with 93% in the same-race tasks), but became more willing with a \$1 incentive (85%), and perfectly willing with a \$5 incentive (100%). Unlike in the same-race judgments, therefore, incentives had a significant impact on choice, $\chi^2(2) = 8.88, p < .02$ (Figure 2).

Thus some 15% of participants were willing to forgo a chance at \$1 rather than make a choice between a White and a Black face (while just 5% did so with two White faces), demonstrating a willingness to incur costs to appear racially neutral. This desire was overwhelmed when choosing had an expected value of \$2.50, however, at which point nearly every participant was willing to choose.

Once again, we examined which face was chosen among participants who did make a choice. The percentage of participants making a choice was higher than in Study 1 due to the incentives manipulation; participants were more likely to select the Black (53%) than White (30%) candidate.

Study 3: Neural Mechanisms of Racial Paralysis

Results of the first two studies suggest that people are highly motivated to forego choices that threaten normative prescriptions for racial neutrality. Our first goal in Study 3 was to identify moderating factors and potential boundary conditions of racial paralysis. In particular, we explored whether all cross-race choices increase opting out, or if this response is specific to cross-race choices involving traits that are relevant to the Black stereotypes (e.g., “intelligence”). We suggest that opting out is a strategic response that people adopt to avoid seeming racially

biased, and therefore predicted that it would be employed more frequently when cross-race choices involved a stereotype-relevant trait.

Thus far, we have not provided direct evidence that cross-race comparisons generate conflict and that opting out is a strategic response that participants settle on after effortful deliberation. Our second goal in Study 3, therefore, was to gain a more complete understanding of the mechanisms underlying racial paralysis by measuring brain activity while participants engaged in a series of same-race and cross-race judgments. In particular, we wished to demonstrate that cross-race decisions were associated with the recruitment of brain regions that detect and signal conflict as well as brain regions that mediate deliberative processing, a finding which would support our contention that cross-race decisions evoke both feelings of uncertainty and concerns about seeming racially biased, and compel people to strategize on how to respond in a socially-appropriate manner.

We therefore expected cross-race judgments to be associated with recruitment of the anterior cingulate cortex (ACC), which monitors for conflict and signals the need for controlled processing and further deliberation (Botvinick, Braver, Barch, Carter, & Cohen, 2001; Carter et al., 1998; Kerns et al., 2004; Lieberman, 2003; 2007; Petersen, Fox, Posner, Mintun, & Raichle, 1988). Second, we expected greater activity in dorsolateral prefrontal cortex (DLPFC), which supports efforts to collect, deliberate on, and integrate information before choosing (Christoff & Gabrieli, 2000; Goel & Dolan, 2000; Waltz et al., 1999) during cross-race choices. Third, we assessed activity in ventromedial prefrontal cortex (VMPFC), a brain region implicated in encoding and signaling the emotional value of decisions and behaviors, especially those that threaten normative and moral prescriptions (Beer, Heerey, Keltner, Scabini, & Knight, 2003; Camille et al., 2004; Damasio, Tranel, & Damasio, 1991; Greene, Sommerville, Nystrom, Darley, & Cohen, 2001; Koenigs et al., 2007; Krajchich, Adolphs, Tranel, Denburg, & Camerer,

2009). We expected that activity in the VMPFC during cross-race judgments would depend on the relevance of the trait in question to Black stereotypes, such that cross-race decisions involving stereotype-relevant traits would be associated with significantly greater VMPFC activity. Finally, we sought evidence that cross-race choices were associated with increased recruitment of the ventrolateral prefrontal cortex (VLPFC), a region implicated in inhibiting preferred but contextually inappropriate responses (Casey et al., 1997; Kowalska, Bachevalier, & Mishkin, 1991).

After first establishing the hypothesized behavioral effect – increased opting out of cross-race choices involving traits that are relevant to Black stereotypes – we measured cortical activity while participants made cross-race and same-race choices in a magnetic resonance imaging (MRI) scanner.

Pilot Behavioral Study

Method

Participants ($N = 46$; 36 Asian, 8 White, 1 Native American, 1 Hispanic) participated in exchange for monetary compensation. The experiment had a 2 (choice set: same-race, cross-race) X 2 (stereotype relevance: relevant, irrelevant) repeated-measures design.

Upon arrival in the laboratory, each participant was greeted by a female experimenter and directed to sit in front of a Dell PC computer. Participants were informed that they would see two faces on the screen at a time with a single characteristic listed at the top of the screen, and that their task was to indicate, via a key press, which person was more likely to exemplify the characteristic that was listed or whether they had no gut feeling. For each trial, a fixation-cross appeared at the center of the screen for 3000 ms, then was replaced with a screen displaying two

faces side by side and the phrase “I have no gut feeling” between the faces. The target trait appeared at the top of the screen. This screen remained visible until a response was recorded; and participants completed a total of 90 trials.

Stimuli comprised a total of 60 different male faces presented on a black background. Fifteen of the images presented Black males, and 45 images presented White males, such that there were 15 cross-race (one White and one Black) and 15 same-race (two White) choices. Each image was approximately 150 x 200 pixels. A total of 30 different characteristics were used in the study, half of which were relevant to Black stereotypes (e.g., intelligent, articulate) and half of which were irrelevant (e.g., restless, strict) based on pre-testing (Appendix A).

Results and Discussion

A 2 (choice set: same race, cross-race) x 2 (stereotype relevance: relevant, irrelevant) repeated measures analysis of variance (ANOVA) revealed that participants were significantly less likely to make cross-race choices ($M = .79$, $s.e. = .03$) than same-race choices ($M = .82$, $s.e. = .02$), $F(1, 45) = 4.41$, $p = .04$, and were significantly less likely to make choices involving stereotype relevant ($M = .78$, $s.e. = .03$) than irrelevant traits ($M = .83$, $s.e. = .02$), $F(1, 45) = 11.75$, $p = .001$.² These effects were qualified by a marginally significant interaction, $F(1, 45) = 3.23$, $p < .08$, which as predicted was driven by the fact that opt out rates were significantly higher when participants made cross-race decisions about stereotype-relevant traits relative to when they made choices in the other three contexts, $F(1, 45) = 4.50$, $p < .04$. Consistent with our hypothesis, participants were significantly less likely to make choices involving relevant ($M = .76$, $s.e. = .03$) than irrelevant traits ($M = .83$, $s.e. = .02$) when making cross-race choices, $t(45) = 3.57$, $p = .001$, while choice rates for same-race did not depend on the relevance of the trait, $t(45) < 1$, ns (Figure 3).

Imaging Study

Having established behaviorally our prediction that participants were most likely to opt out of cross-race judgments involving characteristics relevant to Black stereotypes, we next conducted an imaging study – using the same 2 (choice set: same-race, cross-race) x 2 (stereotype relevance: relevant, irrelevant) repeated measures design – to explore brain regions associated with racial paralysis.

Method

Participants ($N = 18$; 12 females; mean age = 22.7; 9 Caucasians, 2 Hispanics, 7 Asians) completed the experiment for monetary compensation. All participants were strongly right-handed as measured by the Edinburgh handedness inventory (Raczowski, Kalat, & Nebes, 1974), reported no significant abnormal neurological history, and had normal or corrected-to-normal visual acuity.

Stimuli comprised the same 60 faces and 20 of the traits – ten relevant, ten irrelevant – used in the pilot behavioral study (Appendix A). Participants were given the same instructions as in the behavioral pilot – that they would see two faces on the screen and a trait and that their task was to indicate via response keys which person was more likely to exemplify the characteristic, or to indicate that they had no gut feeling. Each trial had the same format as in the behavioral study; participants in the imaging study completed a total of 120 trials.

Participants were scanned in two event-related functional (EPI) runs. A total of 147 volumes were collected in each EPI run. Across the two runs, participants completed 30 of each trial type for a total of 120 trials. Each trial lasted for a duration of 1.5 TRs (the TR was 2 seconds). The remaining 57 EPI volumes were jittered catch trials (i.e., fixation symbols, “+”) used to optimize estimation of the event-related BOLD response. The stimuli were presented using Presentation (version 12.1) and back projected with an LCD projector onto a screen at the

end of the magnet bore that participants viewed by way of a mirror mounted on the head coil. Pillow and foam cushions were placed within the head coil to minimize head movements. All images were collected using a GE scanner with standard head coil. T1- weighted anatomical images were collected using a 3-D sequence (SPGR; 180 axial slices, TR = 19 ms, TE = 5 ms, flip angle = 20°, FOV = 25.6 cm, slice thickness = 1 mm, matrix = 256 x 256). Functional images were collected with a gradient echo EPI sequence (each volume comprised 27 slices; 4 mm thick, 0 mm skip; TR = 2000 ms, TE = 35 ms, FOV = 19.2 cm, 64 x 64 matrix; 84° flip angle).

fMRI Analysis. Functional MRI data were analyzed using Statistical Parametric Mapping software (SPM5, Wellcome Department of Cognitive Neurology, London, UK; Friston et al., 1995). For each functional run, data were preprocessed to remove sources of noise and artifact. Preprocessing included slice timing and motion correction, coregistration to each participant's anatomical data, normalization to the ICBM 152 brain template (Montreal Neurological Institute), and spatial smoothing with an 8 mm (full-width-at-half-maximum) Gaussian kernel. Analyses took place at two levels: formation of statistical images and regional analysis of hemodynamic responses. For each participant, a general linear model with 30 regressors was specified. For each run, the model included regressors specifying the four conditions of interest (modeled with functions for the hemodynamic response), six motion-related regressors, a regressor for each of the first four brain volumes collected, and a regressor constant term that SPM automatically generates and includes in the model. The general linear model was used to compute parameter estimates (β) and t-contrast images for each comparison at each voxel. These individual contrast images were then submitted to a second-level, random-effects analysis to obtain mean t-images.

Results and Discussion

As with the analyses from the behavioral pilot, we first examined overall differences between cross-race and same-race choices. To determine which regions were more active when participants made cross-race relative to same-race choices, regardless of the stereotype-relevance of the trait, we computed the direct contrast '*cross-race choices > same-race choices*', $p < .001$; $k = 10$. Consistent with the view that cross-race choices evoke conflict and feelings of uncertainty, the ACC (-6 33 24; BA32) was significantly more active during cross-race relative to same-race choices. Cross-race choices were also associated with greater recruitment of bilateral DLPFC (-15 42 31; 27 51 23; BA9), a brain area that supports explicit attempts by decision-makers to reflect, integrate and deliberate on information before choosing, and greater recruitment of bilateral VLPFC (-33 26 -11; 36 20 -18; BA47), a brain area that plays a central role in inhibiting instinctively preferred but contextually inappropriate responses. No brain regions exhibited significantly greater activity while participants made same relative to cross-race choices at this threshold (Table 1; Figure 3).

We next explored regions which tracked with choices involving stereotype-relevant versus irrelevant characteristics, regardless of whether the choice was same-race or cross-race. We computed the direct contrast '*relevant > irrelevant*', $p < .001$; $k = 10$. A bilateral cluster in the ACC (9 25 36; BA32), a cluster that extended across the left putamen into the insula (-33, 14, -7), and a small cluster in the right posterior gyrus (-48, -19, 30; BA2) were significantly more active when participants made judgments about stereotype relevant relative to irrelevant traits. No brain regions exhibited significantly greater activity while participants made decisions about irrelevant relative to relevant traits at this threshold.

Finally, we examined regions which were significantly more active during cross-race choices about relevant traits – the judgments we predicted would be most likely to elicit racial paralysis, and confirmed by the behavioral data – relative to the other three judgments.

Consistent with our predictions, cross-race choices about relevant traits were uniquely associated with recruitment of the VMPFC (0 50 -6; BA10), $p < .001$; $k = 10$, a brain area that plays a central role in signaling the emotional value of decisions and behaviors (Figure 4). Furthermore, stereotype relevance moderated the effect of choice set in the ACC (-12 27 21; BA32), bilateral DLPFC (30 19 32; -18 45 27; BA9), and bilateral VLPFC (-27 20 -14; 33 20 -14; BA47). No brain regions exhibited significantly less activity during cross-race choices about relevant traits compared to the other three choice contexts at this threshold (Table 2).

Study 3 demonstrates a role for a key moderator of the tendency to opt out of cross-race decisions: the relevance of the particular decision to stereotypes about Black Americans. As we expected, and the behavioral data confirm, opting out of cross-race decisions was more pronounced for more sensitive judgments than more innocuous judgments. This increased sensitivity to stereotype-relevant judgments was accompanied by increased VMPFC recruitment, a brain region implicated in self-conscious emotions that plays a central role in the regulation of behaviors and judgments governed by strong social and moral norms. In addition, cross-race decisions – when compared with same-race decisions – were associated with increased activation of ACC, DLPFC, and VLPFC, regions involved with conflict, deliberation, and inhibition of responses, respectively. The implication of these regions in cross-race decisions offers support for our account that the fear of appearing biased evoked by such situations leads to conflict, greater reflection, and a resulting tendency to opt out.

Study 4: The Effectiveness of Racial Neutrality

The studies thus far have revealed a motivation to opt out of interracial decisions, one which requires sufficient incentives to overcome and is rooted in regions of the brain involved in conflict, deliberation, and inhibition. In Study 4, we assessed whether these costs of opting out

have some payoff, by exploring whether observers of such behavior do see opting out as an effective means of appearing unbiased. Most importantly, we assessed perceptions of effectiveness among Whites, but also among Blacks; in addition, we moved from using primarily college-aged samples to a more representative national sample, to increase the generalizability of our investigation. We noted in the Introduction that while Whites may be particularly concerned with appearing racially biased, Blacks are not immune from such concerns; while a White person who chooses a Black person might be seen as racist, a Black person who chooses a Black person might be seen as biased in favor of his own group. In Study 4, therefore, we also asked both White and Black respondents to engage in the *Valedictorian* and *Criminal* tasks we used in Study 1; we expected both Whites and Blacks to opt out of cross-race decisions.

In short, despite the fact that colorblindness has been shown to negatively impact interracial interactions while norms of multiculturalism have been shown to lead to less biased attitudes and improved intergroup interactions (e.g., Vorauer, Gagnon, & Sasaki, 2009; Wolsko et al., 2000), we expected both Black and White observers to opt out of cross-race decisions, and to rate others who opted out of cross-race decisions as less biased – precisely because racial paralysis is seen as an effective means of demonstrating racial neutrality.

Method

Respondents ($N = 296$) were recruited by an online survey research company and paid \$5 for their participation. They were selected from a panel of 2.5 million respondents matched to the most recent United States Census on age, education level, and median income. We oversampled on Black respondents in order to have equal numbers of White ($N = 151$) and Black ($N = 145$) respondents (see Table 3 for demographics).

Respondents were assigned to either the *Valedictorian* or *Criminal* tasks from Study 1; in Study 4, all respondents made cross-race choices (between a White and Black candidate). We told respondents we had asked other respondents from the same sample to make these choices (“Which person do you think you will have more success in college?” or “Which person do you think spent time in jail for violent assault?”) and asked respondents to rate how biased they thought people were who made each selection (i.e., chose the White person, the Black person, or chose “no gut feeling,”) on a 5-point scale (1 = *not at all* to 5 = *very*).

Finally, we asked respondents to indicate what choice they would make, by selecting the White candidate, the Black candidate, or “no gut feeling.”

Results and Discussion

Ratings. As we expected, respondents rated people who selected the no gut feeling option as less biased than others who selected either candidate, across both versions of the task – and this was true for both White and Black respondents.

In the *Valedictorian* task, respondents rated people who chose the White candidate ($M = 3.16$, $SD = 1.18$) or the Black candidate ($M = 2.88$, $SD = 1.19$) as more biased than those who opted out ($M = 2.49$, $SD = 1.23$), $F(2, 428) = 24.32$, $p < .001$. While respondents rated people who chose the White candidate as more biased than those who chose the Black candidate, $t(215) = 3.57$, $p < .001$, ratings of bias of people who chose either candidate were significantly higher than ratings of people who opted out, $ts(215) > 3.90$, $ps < .001$. There was no main effect of respondent race, $F < 1$, and no interaction, $F(2, 428) = 2.18$, $p > .11$.

In the *Criminal* task, respondents again rated people who chose either the White candidate ($M = 2.97$, $SD = 1.03$) or the Black candidate ($M = 3.43$, $SD = 1.18$) as more biased than those who opted out ($M = 2.62$, $SD = 1.24$), $F(2, 156) = 11.69$, $p < .001$. Respondents rated people who chose the Black candidate as more biased than those who chose the White candidate,

$t(79) = 2.87, p < .01$; most importantly for our account, however, ratings of bias for people for choosing either candidate were again significantly higher than ratings for opting out, $ts(79) > 2.15, ps < .04$. There was a main effect of respondent race, $F(1,78) = 5.93, p < .02$, such that Black respondents' ratings of bias were on average higher than White respondents' ratings, but again no interaction with respondent race, $F < 1$.

The lack of interaction for both versions of the task demonstrates that Black and White respondents agreed that people who opted out were less biased than people who made a choice. Figures 5A and 5B demonstrate the consensus between Black and White respondents.

Choice. As in the previous studies, respondents' own choices also reflected a strong tendency to opt out of cross-race choices. For the *Valedictorian* task, overall some 56% chose no gut feeling, with just 25% choosing the White candidate, and 19% choosing the Black candidate, $\chi^2(2) = 50.86, p < .001$; these results were strikingly similar for Black and White respondents, with both Blacks (57%) and Whites (55%) selecting "no gut feeling" the majority of the time, $\chi^2s(2) > 21.88, ps < .001$. Similarly for the *Criminal* task, fully 75% of respondents chose no gut feeling, with just 17% choosing the White candidate, and 8% choosing the Black candidate, $\chi^2(2) = 63.70, p < .001$; these results were again similar for Black and White respondents, with both Blacks (65%) and Whites (82%) selecting "no gut feeling" the majority of the time, $\chi^2s(2) > 14.00, ps < .001$.

Finally, ratings of the bias indicated by choosing was related to respondents' own choices – the correlation between respondents' rating of another person's choice of "no gut feeling" was significantly correlated with respondents' own tendency to select this option, $r(296) = -.20, p < .01$. The more that respondents perceived others who opted out as unbiased, the more likely they were to opt out themselves.

General Discussion

Four studies demonstrate that while people are willing to make choices between two members of the same group (e.g., two White males, two Black males) on the basis of nothing more than their photographs, they experience racial paralysis when making judgments about members of different groups (a White and a Black male), choosing to opt out of such decisions altogether. Somewhat ironically, people's efforts to honor racial neutrality by not choosing provides the very evidence that they do notice race; after all, if they truly didn't notice race, they would be as likely to make choices in same-race and cross-race judgments. This tendency to opt out held across choices in different domains and with different valences, from choosing a valedictorian to choosing a criminal (Study 1). The tendency was more pronounced, however, for judgments that were more relevant to stereotypes about Blacks – and therefore more likely to elicit concerns about appearing biased – as reflected both in opt out rates and activation in brain regions related to emotionally guided choice (Study 3). Thus despite the extraordinarily and variegated abilities of humans to decode faces in order to facilitate judgments and decisions about others, changing the context seems to abruptly change these processes: Absent some strong incentive to act (such as the \$5 incentive in Study 2), Whites exhibit racial paralysis. Finally, Study 4 demonstrates that coping with the stress of cross-race decisions by opting out may in fact be the wisest path to appearing unbiased, as both White and Black Americans agree that decision makers who opt out – and honor racial neutrality – have demonstrated a lack of bias.

This is not to suggest that this reluctance is universal across all individuals and all choices. First, while individuals across the political spectrum are motivated to appear unbiased – reporting more positive attitudes than implicit measures reveal (Nosek et al., 2002) – people's motivation to appear unbiased (Dunton & Fazio, 1997; Plant & Devine, 1998) may predict Whites' avoidance of choice. Second, situational factors, such as making choices more public or

assuaging people's concern about appearing biased (Monin & Miller, 2001) would likely moderate our results. Finally, we have focused on the most salient judgment, between one White male and one Black male, but the judgment tasks we use here could incorporate other ethnicities or social categories (e.g., gender, or physical disability) to explore more generally the unwillingness to make choices between members of different social groups; indeed, the frequency with which people opt out of decisions between members of different social groups (for example, between an obese and normal weight person) could be used a metric for concerns about appearing biased towards those groups (Crandall, Eshleman, & O'Brien, 2002).

Our task, which focuses on simple judgments between two faces of members of two different social groups, is a clear abstraction from the kinds of situations with which we opened the paper: choosing whom to sit next to on the subway, or talk to in an elevator. As we noted at the beginning, these innocuous real-world situations are in themselves less serious instantiations of more consequential real-world decisions, such as who to hire, admit to college, or send to jail. Results from Study 3 demonstrate that as the stakes of some choice get higher (when making a choice feels more likely to indicate that one is biased) the incidence of racial paralysis increases. Our proxy for importance, of course, was the relevance of the judgment to some stereotype about Blacks. We can only imagine the racial paralysis that might ensue during discussions of real-world impactful decisions, where speaking in favor of a White candidate over a Black candidate makes one appear racist, whereas speaking in favor of a Black candidate over a White candidate can make one appear as though one is trying too hard *not* to appear racist. We suspect that in such discussions, when people are forced to make some decision at the end of the day, decision makers rely on other strategies to avoid the appearance of bias, such as deferring to members of minority groups (Crosby, Monin, & Richardson, 2008).

Conclusion

One view of Whites' tendency to opt out of cross-race decisions is that such behavior constitutes an error, echoing other research suggesting that efforts to appear racially neutral can impede both performance and intergroup interactions (Norton et al., 2006; Richeson & Nussbaum, 2004). By this standard, the normative behavior in our paradigm would be for 100% of people to select a face; indeed, people approach this normative standard given sufficient monetary incentives. At the societal level, however, the normative standard is less clear. While people occasionally may be willing to allow others to judge them on the basis of scant information, such as fortune tellers providing a life history on the basis of one's date of birth, most would not be pleased to be judged solely on the basis of their physical appearance. The fact that people are able and willing to make inferences about others on the basis of their faces – as in the present study – or mere snippets of observation (Ambady, Bernieri, & Richeson, 2000; Mehl, Gosling, & Pennebaker, 2006) does not necessarily mean that they should do so. Indeed, in some sense the exact opposite behavior might be more desirable: People should be *unwilling* to judge others on the basis of such factors. Thus though racial neutrality may inhibit behavior, the wisdom or folly of this behavior ultimately depends on which standard one deems most important. While often seen as unnecessarily inhibitory, racial neutrality may serve as a general reminder of the broader principle that people of all social groups should be judged by qualities other than their physical appearance.

References

- Ambady, N., Bernieri, F. J., & Richeson, J. A. (2000). Toward a histology of social behavior: Judgmental accuracy from thin slices of the behavioral stream. In M. P. Zanna (Ed.), *Advances in Experimental Social Psychology*, 32 (pp. 201-272). San Diego, CA: Academic Press.
- Apfelbaum, E. P., Pauker, K., Ambady, N., Sommers, S. R., & Norton, M. I. (2008). Learning (not) to talk about race: When older children underperform in social categorization. *Developmental Psychology*, 44, 1513-1518.
- Apfelbaum, E. P., Sommers, S. R., & Norton, M. I. (2008). Seeing race and seeming racist? Evaluating strategic colorblindness in social interaction. *Journal of Personality and Social Psychology*, 95, 918-932.
- Baron, R. S., Vandello, J. A., & Brunsman, B. (1996). The forgotten variable in conformity research: The impact of task importance on social influence. *Journal of Personality and Social Psychology*, 71, 915-927.
- Beer, J. S., Heerey, E. H., Keltner, D., Scabini, D., & Knight, R. T. (2003). The regulatory function of self-conscious emotion: Insights from patients with orbitofrontal damage. *Journal of Personality and Social Psychology*, 85, 594-604.
- Blair, I. V., Judd, C. M., Sadler, M. S., & Jenkins, C. (2002). The role of Afrocentric features in person perception: Judging by features and categories. *Journal of Personality and Social Psychology*, 83, 5-25.
- Bonilla-Silva, E. (2003). *Racism without racists: Color-blind racism and the persistence of racial inequality in the United States*. Lanham, MD: Rowman & Littlefield
- Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., & Cohen, J. D. (2001). Conflict monitoring and cognitive control. *Psychological Review*, 108, 624-652.

- Brett, M., Anton, J.L., Valabregue, R., & Poline, J.B. (2002). Region of interest analysis using an SPM toolbox. *NeuroImage*, 16, abstract 497 (available on CD-ROM).
- Camille, N., Coricelli, G., Sallet, J., Paradat-Diehl, P., Duhamel, J. R., & Sirigu, A. (2004). The involvement of the orbitofrontal cortex in the experience of regret. *Science*, 304, 1167-1170.
- Carbado, D. W., & Harris, C. I. (2008). The new racial preferences. *California Law Review*, 96, 1139-1214.
- Carter, C. S., Braver, T. S., Barch, D. M., Botvinick, M. M., Noll, D. C., Cohen, J. D. (1998). Anterior cingulate cortex, error detection, and the online monitoring of performance. *Science*, 280, 747-749.
- Casey, B.J., Castellanos, F.X., Giedd, J.N., Marsh, W.L., Hamburger, S.D., Schubert, A.B., Vauss, Y.C., Vaituzis, A.C., Dickstein, D.P., Sarfatti, S.E., & Rapoport, J.L. (1997). Implication of right frontostriatal circuitry in response inhibition and attention-deficit/hyperactivity disorder. *Journal of the American Academy of Child and Adolescent Psychiatry*, 36, 374–383
- Chernoff, H. (1973). Use of faces to represent points in k-dimensional space graphically. *Journal of the American Statistical Association*, 68, 361-368.
- Christoff, K. & Gabrieli, J.D.E. (2000). The frontopolar cortex and human cognition: Evidence for a rostrocaudal hierarchical organization within the human prefrontal cortex. *Psychobiology*, 28, 168-186.
- Correll, J., Park, B., Judd, C. M., & Wittenbrink, B. (2002). The police officer's dilemma: Using ethnicity to disambiguate potentially threatening individuals. *Journal of Personality and Social Psychology*, 83, 1314–1329.

- Crandall, C. S., & Eshelman, A. (2003). A justification-suppression model of the expression and experience of prejudice. *Psychological Bulletin*, 129, 414–446.
- Crandall, C. S., Eshelman, A., & O'Brien, L. T. (2002). Social norms and the expression and suppression of prejudice: The struggle for internalization. *Journal of Personality and Social Psychology*, 82, 359-378.
- Crosby, F., Bromley, S., & Saxe, L. (1980). Recent unobtrusive studies of Black and White discrimination and prejudice: A literature review. *Psychological Bulletin*, 87, 546–563.
- Crosby, J. R., Monin, B., & Richardson, D. (2008). Where do we look during potentially offensive behavior? *Psychological Science*, 19, 226-228.
- Cuddy, A. J. C., Fiske, S. T., & Glick, P. (2008). Warmth and competence as universal dimensions of social perception: The Stereotype Content Model and the BIAS Map. In M. P. Zanna (Ed.), *Advances in Experimental Social Psychology* (vol. 40, pp. 61-149). New York, NY: Academic Press.
- Damasio, A. R., Tranel, D., & Damasio, H. (1991). Somatic markers and the guidance of behavior: Theory and preliminary testing. In H.S. Levin, H.M. Eisenberg, & A.L. Benton (Eds.), *Frontal lobe function and dysfunction* (pp. 217-229). New York: Oxford University Press.
- Devine, P. G. (1989). Stereotypes and prejudice: Their automatic and controlled components. *Journal of Personality and Social Psychology*, 56, 5–18.
- Devine, P. G., & Elliot, A. J. (1995). Are racial stereotypes really fading? The Princeton Trilogy revisited. *Personality and Social Psychology Bulletin*, 21, 1139–1150.
- Dhar, R. & Simonson, I. (2003). The effect of forced choice on choice. *Journal of Marketing Research*, 40, 146-150.

- Dovidio, J. F. & Gaertner, S. L. (1981). The effects of race, status, and ability on helping behavior. *Social Psychology Quarterly*, 44, 192-203.
- Dovidio, J. F., & Gaertner, S. L. (2004). Aversive racism. In M. P. Zanna (Ed.), *Advances in Experimental Social Psychology* (vol. 36, pp. 1-51). San Diego, CA: Academic Press.
- Dovidio, J. F., Kawakami, K., & Gaertner, S. L. (2002). Implicit and explicit prejudice and interracial interaction. *Journal of Personality and Social Psychology*, 82, 62–68.
- Dunton, B. C. & Fazio, R. H. (1997). An individual difference measure of motivation to control prejudiced reactions. *Personality and Social Psychology Bulletin*, 23, 316-326.
- Fazio, R. H., Jackson, J. R., Dunton, B. C., & Williams, C. J. (1995). Variability in automatic activation as an unobtrusive measure of racial attitudes: A bona fide pipeline? *Journal of Personality and Social Psychology*, 69, 1013–1027.
- Friston, K. J., Holmes, A. P., Worsley, K. J., Poline, J. B., Frith, C. D., & Frackowiack, R.J.S. (1995). Statistical parametric maps in functional imaging: A general linear approach. *Human Brain Mapping*, 2, 189-210.
- Gaertner, S. L. & Dovidio, J. F. (1986). The aversive form of racism. In J. F. Dovidio, & S. L. Gaertner (Eds.), *Prejudice, discrimination, and racism* (pp. 61-89). Orlando, FL: Academic Press.
- Goel, V. & Dolan, R. J. (2000). Anatomical segregation of component processes in an inductive inference task. *Journal of Cognitive Neuroscience*, 12, 110-119.
- Greene, J. D., Sommerville, R. D., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science*, 293, 2105-2108.
- Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review*, 102(1), 4-27.

- Greenwald, A. G., McGhee, D. E., & Schwarz, J. L. K. (1998). Measuring individual differences in implicit cognition: The Implicit Association Test. *Journal of Personality and Social Psychology*, 74(6), 1464-1480.
- Hassin, R. & Trope, Y. (2000). Facing faces: Studies on the cognitive aspects of physiognomy. *Journal of Personality and Social Psychology*, 78, 837-852.
- Hodson, G., Dovidio, J. F., & Gaertner, S. L. (2002). Processes in racial discrimination: Differential weighting of conflicting information. *Personality and Social Psychology Bulletin*, 28, 460-471.
- Ito, T. A., & Urland, G. R. (2003). Race and gender on the brain: Electrocortical measures of attention to the race and gender of multiply categorizable individuals. *Journal of Personality and Social Psychology*, 85, 616-626.
- Johansson, P., Hall, L., Sikstrom, S., & Olsson, A. (2005). Failure to detect mismatches between intention and outcome in a simple decision task. *Science*, 310, 116-119.
- Jost, J. T., Rudman, L. A., Blair, I. V., Carney, D. R., Dasgupta, N., Glaser, J., & Hardin, C. D. (2009). The existence of implicit bias is beyond scientific doubt: A refutation of ideological and methodological objections and executive summary of ten studies that no manager should ignore. *Research in Organizational Behavior*, 29, 39-69.
- Katz, D., & Braly, K. W. (1933). Racial stereotypes of one-hundred college students. *Journal of Abnormal and Social Psychology*, 28, 280-290.
- Kerns, J. G., Cohen, J. D., MacDonald, A. W., Cho, R. Y., Stenger, V. A., & Carter, C. S. (2004). Anterior cingulate conflict monitoring and adjustments in control. *Science*, 303, 1023-1026.

- Koenigs, M., Young, L., Adolphs, R., Tranel, D., Cushman, F., Hauser, M., & Damasio, A. (2007). Damage to the prefrontal cortex increases utilitarian moral judgments. *Nature*, 446, 908-911.
- Kowalska, D.M., Bachevalier, J., & Mishkin, M. (1991). The role of the inferior prefrontal convexity in performance of delayed nonmatching-to-sample. *Neuropsychologia*, 29, 583– 600.
- Krajchich, I., Adolphs, R., Tranel, D., Denburg, N. L., & Camerer, C. F. (2009). Economic games quantify diminished sense of guilt in patients with damage to the prefrontal cortex. *The Journal of Neuroscience*, 29, 2188-2192.
- Larrick, R. P. (1993). Motivational factors in decision theories: The role of self-protection. *Psychological Bulletin*, 113, 440-450.
- Lieberman, M. D. (2003). Reflective and reflexive judgment processes: A social cognitive neuroscience approach. In J.P. Forgas, K.R. Williams, & W. von Hippel (Eds.), *Social judgments: Implicit and explicit processes* (pp. 44-67). New York: Cambridge University Press.
- Lieberman, M. D. (2007). Social cognitive neuroscience: A review of core processes. *Annual Review of Psychology*, 58, 259-289.
- Malpass, R. S. & Kravitz, J. (1969). Recognition for faces of own- and other-race faces. *Journal of Personality and Social Psychology*, 13, 330-334.
- McConahay, J. B. (1986). Modern racism, ambivalence, and the modern racism scale. In J. Dovidio, & S. Gaertner (Eds.), *Prejudice, discrimination, and racism* (pp. 91–125). Orlando, FL: Academic Press.

- Mehl, M. R., Gosling, S. D., & Pennebaker, J. W. (2006). Personality in its natural habitat: Manifestations and implicit folk theories of personality in daily life. *Journal of Personality and Social Psychology*, 90, 862-877.
- Monin, B. & Miller, D. T. (2001). Moral credentials and the expression of prejudice. *Journal of Personality and Social Psychology*, 81, 33-43.
- Montepare, J. M. & Opeyo, A. (2002). The relative salience of physiognomic cues in differentiating faces: A methodological tool. *Journal of Nonverbal Behavior*, 26, 43-59.
- Myrdal, G. (1944). *An American dilemma: The Negro problem and modern democracy*. New York: Harper & Row.
- Nisbett, R. E. & Wilson, T. D. (1977). The halo effect: Evidence for unconscious alteration of judgments. *Journal of Personality and Social Psychology*, 35, 250-256.
- Norton, M. I., Sommers, S. R., Apfelbaum, E. P., Pura, N., & Ariely, D. (2006). Colorblindness and interracial interaction: Playing the Political Correctness Game. *Psychological Science*, 17, 949-953.
- Norton, M. I., Vandello, J. A., & Darley, J. M. (2004). Casuistry and social category bias. *Journal of Personality and Social Psychology*, 87, 817-831.
- Nosek, B. A., Banaji, M. R., & Greenwald, A. G. (2002). Harvesting implicit group attitudes and beliefs from a demonstration website. *Group Dynamics*, 6, 101-115.
- Pager, D. & Quillian, L. (2005). Walking the talk? What employers say versus what they do. *American Sociological Review*, 70, 355-380.
- Petersen, S. E., Fox, P. T., Posner, M. I., Mintun, M., & Raichle, M. (1988). Positron emission tomographic studies of the cortical anatomy of single word processing. *Nature*, 331, 585-589.

- Plant, E. A., & Devine, P. G. (1998). Internal and external motivation to respond without prejudice. *Journal of Personality and Social Psychology*, 69, 811-832.
- Plaut, V.C., Thomas, K.M., & Goren, M.J. (2009). Is multiculturalism or colorblindness better for minorities? *Psychological Science*, 20, 444-446.
- Pollock, M. (2004). *Colormute: Race talk dilemmas in an American school*. New Jersey: Princeton University Press.
- Regents of the University of California v. *Bakke*, 438 U.S. 265 (1978).
- Raczkowski, D., Kalat, J. W., & Nebes, R. (1974). Reliability and validity of some handedness questionnaire items. *Neuropsychologia*, 12, 43-47.
- Regents of the University of California v. Bakke* (1978). 438 U.S. 265.
- Ricci v. DeStefano*, 557 U.S. ____ (2009).
- Richeson, J. A. & Nussbaum, R. J. (2004). The impact of multiculturalism versus color-blindness on racial bias. *Journal of Experimental Social Psychology*, 40, 417-423.
- Saucier, D. A., Miller, C. T., & Doucet, N. (2005). Differences in helping Whites and Blacks: A meta-analysis. *Personality and Social Psychology Review*, 9, 2-16.
- Schofield, J. W. (2007). The colorblind perspective in school: Causes and consequences. In J. A. Banks & C. A. McGee Banks (Eds.), *Multicultural education: Issues and perspectives* (pp. 271-295). New York: Wiley.
- Smith, M. L., Cottrell, G. W., Gosselin, F., & Schyns, P. G. (2005). Transmitting and decoding facial expressions. *Psychological Science*, 16, 184-189.
- Snyder, M. L., Kleck, R. E., Strenta, A., & Mentzer, S. J. (1979). Avoidance of the handicapped: An attributional ambiguity analysis. *Journal of Personality and Social Psychology*, 37, 2297-2306.

- Snyder, M., Tanke, E. D., & Berscheid, E. (1977). Social perception and interpersonal behavior: On the self-fulfilling nature of social stereotypes. *Journal of Personality and Social Psychology*, 35, 656–666.
- Solnick S. J. & Schweitzer M. E. (1999). The influence of physical attractiveness and gender on ultimatum game decisions. *Organizational Behavior and Human Decision Processes*, 79, 199-215.
- Valenza, E., Simion, F., Cassia, V. M., & Umiltà, C. (1996). Face preference at birth. *Journal of Experimental Psychology: Human Perception and Performance*, 22, 892-903.
- Vorauer, J. D., Gagnon, A., & Sasaki, S. J. (2009). Salient intergroup ideology and intergroup interaction. *Psychological Science*, 20, 838–845.
- Waltz, J. A., Knowlton, B. J., Holyoak, K. J., Boone, K. B., Mishkin, F. S., de Menezes Santon, M., Thomas, C. R., & Miller, B. L. (1999). A system for relational reasoning in human prefrontal cortex. *Psychological Science*, 10, 119-125.
- Willis, J. & Todorov, A. (2006). First impressions: Making up your mind after 100 ms exposure to a face. *Psychological Science*, 17, 592-598.
- Wolsko, C., Park, B., Judd, C.M., & Wittenbrink, B. (2000). Framing interethnic ideology: Effects of multicultural and color-blind perspectives on judgments of groups and individuals. *Journal of Personality and Social Psychology*, 78, 635–654.
- Zebrowitz, L. A. (1997). *Reading faces: Window to the soul?* Boulder, CO: Westview Press.
- Zebrowitz, L. A. & McDonald, S. (1991). The impact of litigants' babyfacedness and attractiveness on adjudications in small claims courts. *Law and Human Behavior*, 15, 603-623.

Author Note

Michael I. Norton, Harvard Business School; Malia F. Mason, Columbia Business School; Joseph A. Vandello and Andrew Biga, University of South Florida; Rebecca Dyer, Yale University.

The authors thank Dan Ariely, Benoît Monin and Sam Sommers for their helpful suggestions, and Amy Cuddy, Robyn LeBoeuf, Rachel McLoughlin, Sarah Molouki, Katie Offer, Joe Simmons, and Anjuli Wilmer for their assistance with data collection.

Correspondence concerning this article may be addressed to Michael I. Norton (mnorton@hbs.edu), Harvard Business School, Soldiers Field Road, Boston, MA 02163.

Footnotes

1. We used new sets of White and Black faces in Study 2 and additional new sets in Studies 3 and 4 to ensure that our effects were not specific to any one set of faces.
2. Percentage differences were smaller overall in Study 3 when compared with the previous studies. We expected this result given that the repeated-measures nature of the design makes any one decision less consequential. Importantly, however, the predicted effects are significant.

Table 1. Peak coordinates of brain regions that where activity during cross-race choices was significantly greater than activity during same-race choices, $p < .0001$, $k = 10$. The opposite contrast revealed no significant differences. (L.) = Left; (R.) = Right; (BA) = Broadmann Area (Study 3).

A. “cross-race choices > same-race choices”						
k	Anatomical Location	BA	coordinates			t-value
			x	y	z	
30	R. DLPFC	9	27	51	23	5.23
50	L. DLPFC	9	-15	42	31	5.96
113	L. VLPFC	47	-33	26	-11	9.01
40	R. VLPFC	47	36	20	-18	4.75
15	L. ACC	32	-6	33	24	4.64
11	R. superior frontal	6	15	17	54	4.24
40	L. posterior cingulate	30	-18	-61	7	4.43
71	R. cuneus	18	15	-84	15	5.50
119	R. lingual	18	12	-58	7	4.47
14	R. middle temporal	21	60	-29	-5	4.90
11	L. middle temporal	21	-50	-7	-16	4.56
27	R. superior temporal	39	50	-54	25	4.52
11	L. superior temporal	22	-48	11	-4	4.07

Table 2. Peak coordinates of brain regions where activity during cross-race comparisons

involving relevant traits was significantly greater than during other choice contexts $p < .001$, $k =$

10. The opposite contrast revealed no significant differences (B.) = Bilateral; (L.) = Left; (R.) =

Right; (BA) = Broadmann Area (Study 3).

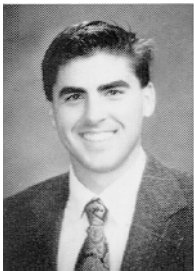

<i>“cross-race/relevant > other three conditions”</i>						
coordinates						
k	Anatomical Location	BA	x	y	z	t-value
35	B. VMPFC	10	0	50	-6	3.77
44	R. DLPFC	9	30	19	32	6.35
95	L. DLPFC	9	-18	45	27	6.32
31	R. DLPFC	9	21	48	27	4.62
44	L. VLPFC	47	-27	20	-14	5.95
44	R. VLPFC	47	33	20	-14	5.52
19	R. VLPFC	47	45	15	-1	5.15
60	L. ACC	32	-12	27	21	4.49
32	R. superior frontal	6	18	11	55	4.66
24	L. superior frontal	6	-9	17	51	4.82
24	R. superior temporal	39	52	-54	21	4.82
10	L. middle temporal	22	57	-41	2	4.41
16	R. hippocampus		24	-38	-2	4.81
10	R. precuneus	31	6	-63	25	3.45
14	R. thalamus		18	-26	1	3.45

Table 3. Demographics by Respondent Race (Study 4)

	White	Black
<hr/>		
Education (%)		
Some High School	.7	1.4
High School Graduate	14.6	17.9
Some College	37.7	37.9
College Graduate	31.3	28.3
Some Graduate School	3.3	4.1
Graduate School Degree	12.6	10.3
<hr/>		
Income (%)		
Less than \$20,000	11.3	12.4
\$20-30,000	8.6	13.1
\$30-40,000	11.9	18.6
\$40-50,000	8.6	14.5
\$50-60,000	13.9	8.3
\$60-70,000	10.6	6.9
\$70-80,000	7.9	9.7
\$80-90,000	5.3	3.4
\$90-100,000	7.3	4.1
More than \$100,000	14.6	9.0
<hr/>		
Gender (%)		
Female	67.5	64.8
Male	32.5	35.2
<hr/>		
Mean Age (SD)	47.2 (12.9)	42.0 (12.2)
<hr/>		

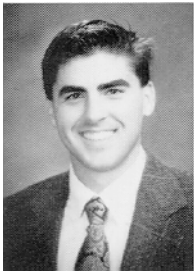
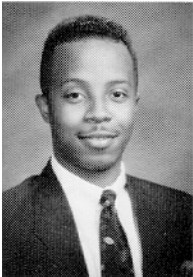
Figure 1. Sample stimuli for the *Criminal* task (Study 1)

Which person do you think spent time in jail for violent assault?

☐ Person A ☐ I have no gut feeling ☐ Person B

Which person do you think spent time in jail for violent assault?

☐ Person A ☐ I have no gut feeling ☐ Person B

Figure 2. Incentives increase the percentage of participants choosing a candidate in cross-race but not same-race choices. Percentages for the same-race choices are averaged across the three filler tasks (Study 2).

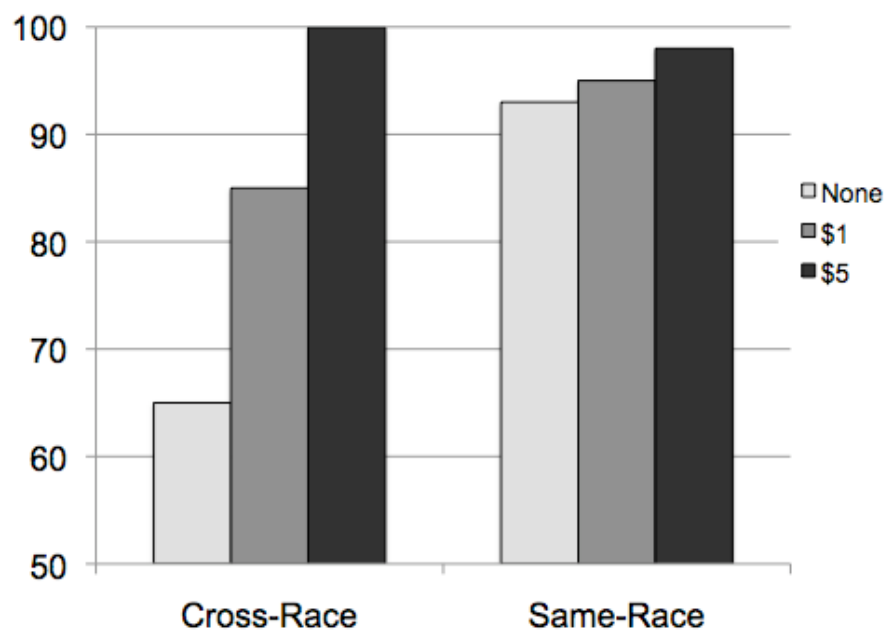


Figure 3. Percent signal change by condition in regions that exhibited a significant main effect of choice set, $p < .001$; $k = 10$. (Top) is a cluster in the ACC (-6 33 24; BA32); (Middle) is a cluster in the right DLPFC (27 51 23; BA9); (Bottom) is a cluster in the left VLPFC (36 20 -18; BA10). Values were computed by dropping a 10 mm sphere at the cluster's peak, extracting the % signal change with the tools provided by the MarsBar interface (Brett, Anton, Valabregue, & Poline, 2002), and then averaging across all participants.

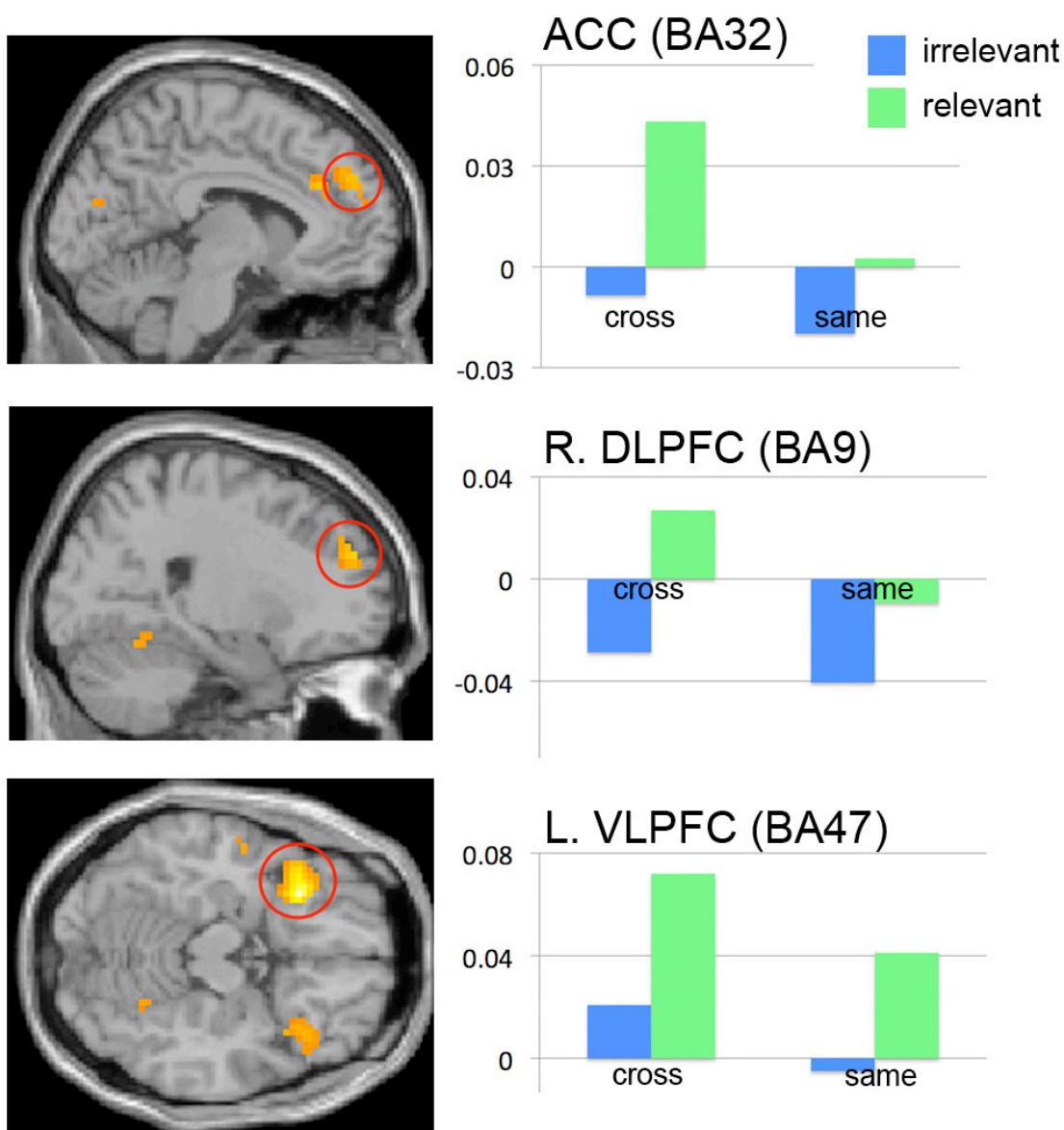


Figure 4. Percent signal change by condition in a region of the VMPFC ($0\ 49\ -2$; BA10) where the activity was significantly greater when participants made cross-race choices about stereotype-relevant traits relative to when participants made choices in the other three decision contexts, $p < .001$; $k = 10$. The values were computed by dropping a 10 mm sphere at the cluster's peak ($-0\ 49\ -2$), extracting the average % signal change with the tools provided by the MarsBar interface (Brett et al., 2002), and then averaging across all participants.

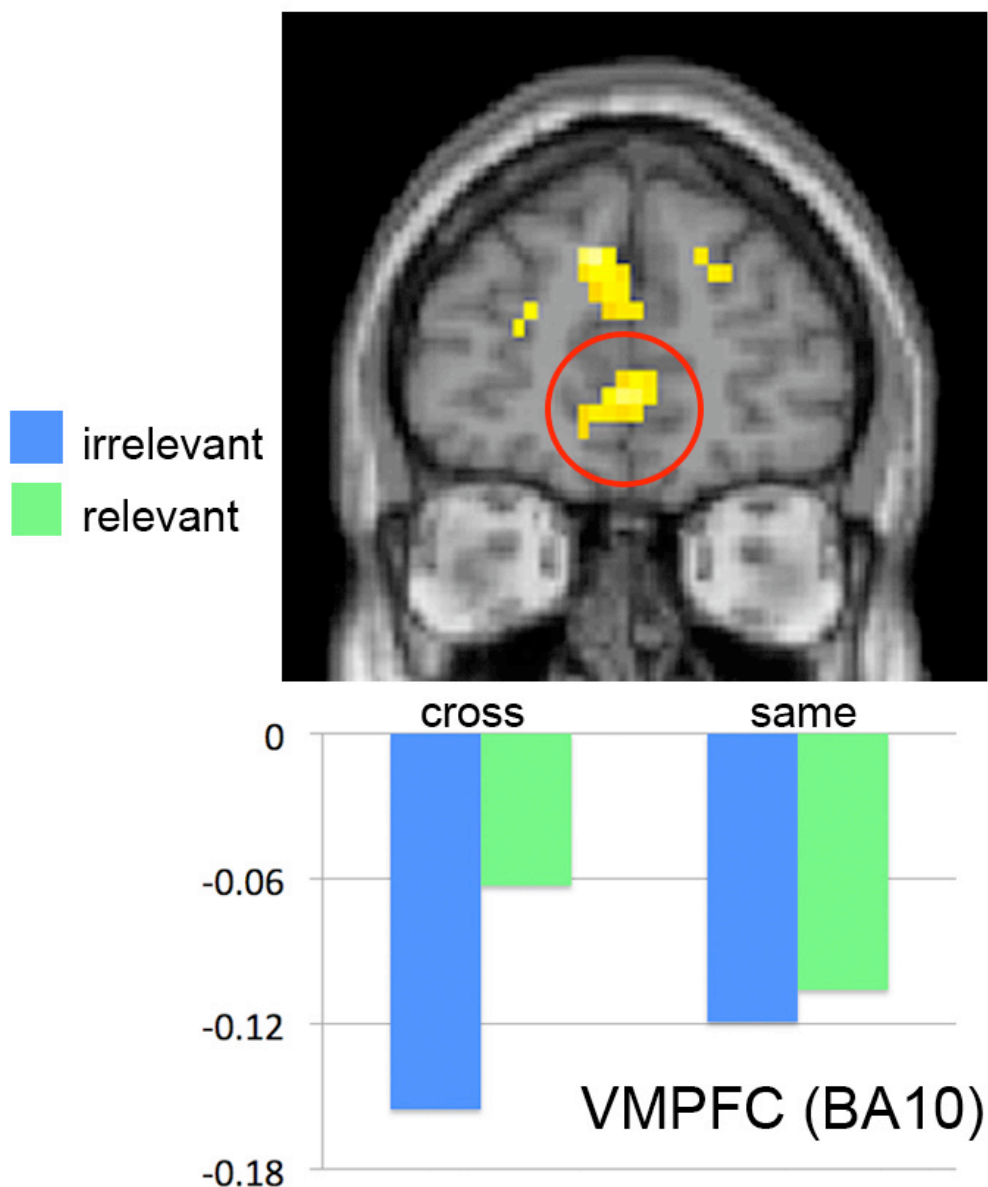
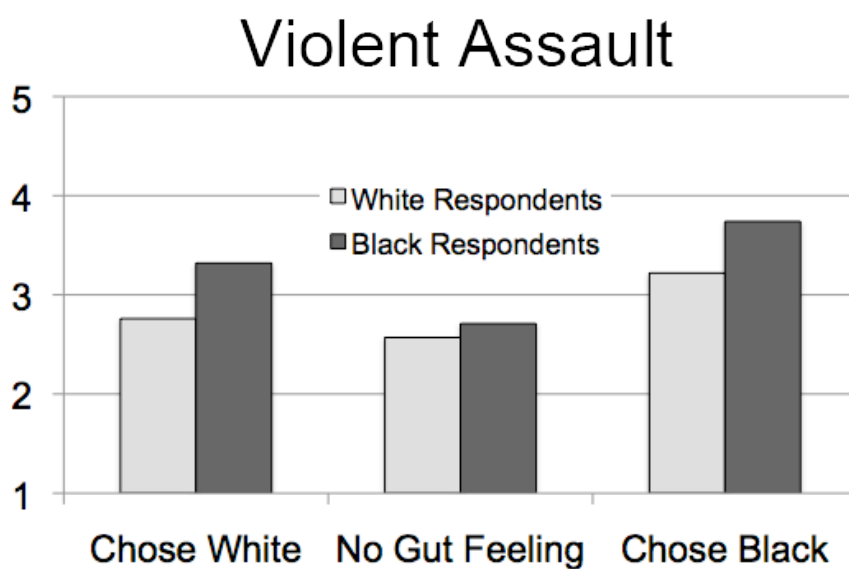
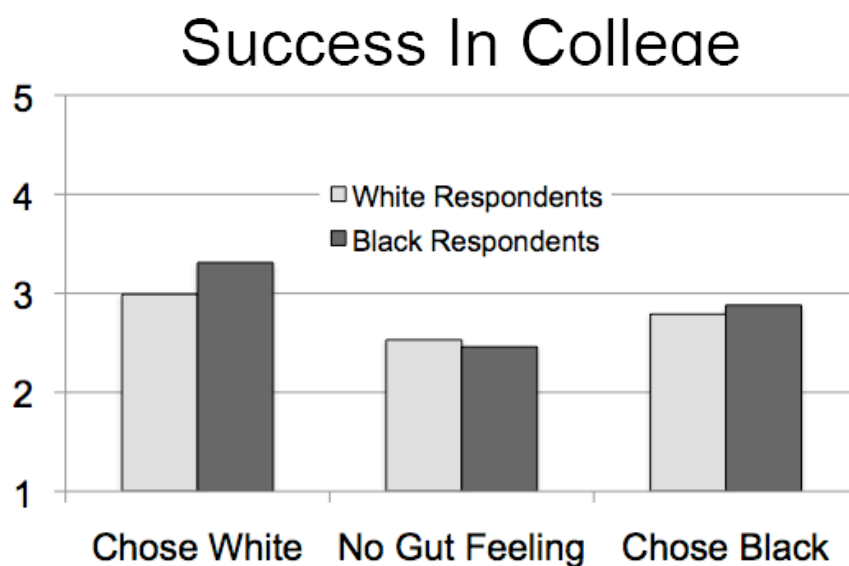


Figure 5A and 5B (Study 4). White and Black respondents agree that choosing “No Gut Feeling” indicates a lower level of bias than choosing either the White or Black candidate, for both versions of the task.



Appendix A**Traits used in Pilot Behavioral Study***Relevant Traits*

- Intelligent
- Motivated
- Articulate
- Responsible
- Competent
- Honest
- Polite
- Agreeable
- Hardworking
- Conscientious
- Reliable
- Patient
- Play golf
- Cultured
- Math major

Irrelevant Traits

- Outgoing
- Quiet
- Restless
- Impressionable
- Strict
- Opinionated
- Loyal
- Self-conscious
- Curious
- Artistic
- Authoritative
- Funny
- From Canada
- Play the guitar
- Have a brother

Traits used in fMRI Study*Relevant Traits*

- Intelligent
- Articulate
- Competent
- Polite
- Agreeable
- Hardworking
- Conscientious
- Reliable
- Patient
- Math major

Irrelevant Traits

- Outgoing
- Restless
- Impressionable
- Strict
- Opinionated
- Loyal
- Curious
- Authoritative
- Play the guitar
- Have a brother