Running head: DISTINGUISHING THREAT VS. NEGATIVE VALENCE AS SOURCES OF AUTOMATIC ANTI-BLACK BIAS

Danger or Dislike:

Distinguishing Threat from Negative Valence as Sources of Automatic Anti-Black Bias

David S. March

Florida State University

Lowell Gaertner & Michael A. Olson

University of Tennessee

In Press, cite as: March, D. S., Gaertner, L., & Olson, M. A. (2021). Danger or Dislike: Distinguishing threat from valence as sources of automatic anti-Black bias. ***Journal of Personality and Social Psychology.***

All of our data, code, survey materials, and supplemental materials are available in the Open Science Framework (OSF) repository for this project (https://osf.io/umyrn/).

Correspondence concerning this article should be addressed to Davis S. March, Department of Psychology, Florida State University, 1107 West Call Street, Tallahassee, FL 32306, United States. Email: march@psy.fsu.edu

Abstract

The Dual Implicit Process Model (March et al., 2018a) distinguishes the implicit processing of physical threat (i.e., “can it hurt or kill me?”) from valence (i.e., “do I dislike/like it?”). Five studies tested whether automatic anti-Black bias is due to White Americans associating Black men with threat, negative valence, or both. Studies 1 and 2 assessed how quickly White participants decided whether positive, negative, and threatening images were good versus bad when primed by Black versus White male-faces. Studies 3 and 4 assessed how early in the decision process White participants began deciding whether Black and White (and, in Study 3, Asian) male-faces displaying anger, sadness, happiness, or no emotion were, in Study 3, dangerous, depressed, cheerful, or calm or, in Study 4, dangerous, negative, or positive. Study 5 assessed how quickly White participants decided whether negative and threatening words were negative versus dangerous when primed by Black versus White male-names. All studies indicated that White Americans automatically associate Black men with physical threat. Study 3 indicated the association is unique to Black men and did not extend to Asian men as a general intergroup effect. Studies 3, 4, and 5, which simultaneously paired threat against negativity, indicated that the Black-threat association is stronger than a Black-negative association.

Keywords: bias, threat, valence, race, implicit

Police Shoot Unarmed Black Man in Florida (New York Times, 2016)

Police Shooting of Unarmed Black Man Roils Sacramento (Wall Street Journal, 2018)

Kansas City Police Shoot, Kill Black Man Officer Thought Was Armed (US News, 2020)

Such headlines are too frequent and familiar, and research indicates that race plays a pivotal role (Correll et al., 2002; Greenwald et al., 2003). As recent events have highlighted, the consequences of anti-Black bias can be deadly. Despite a multitude of ethnicities that could be disliked or negatively stereotyped, police apply force against Blacks 4 times more than against Hispanics and 18 times more than against Asians (and 3 times more than against Whites; Goff et al., 2016). Similarly, when tasked with the decision to "shoot," White participants decide more quickly to shoot armed targets and more slowly to not shoot unarmed targets when those targets are Black rather than Latino, Asian, or White (Sadler et al., 2012). Even Black participants evidence the speeded shooting-bias against Black targets (Correll et al., 2002). Is dislike of Blacks simply stronger than dislike of other stigmatized groups? We suspect not; a more promising approach to understanding these patterns is consideration of the type of threat Black males are perceived to pose (Cottrell & Neuberg, 2005; Neuberg & Schaller, 2016). Despite the range of negative stereotypes with which Blacks are associated, the critical issue might be that White Americans associate Black men with physical threat.

In the current research, we use the Dual Implicit Process Model (DIPM; March et al., 2018a, 2018b) to examine White American's perception of Black men. The DIPM distinguishes the implicit processing of threat ("can it hurt/kill me?") from valence ("do I like/dislike it?"). We apply that distinction to test if White Americans automatically associate Black men with survival threat. Because threatening stimuli are evaluatively negative, we take particular care to methodologically disentangle threat from the more general category of negativity to assess the

possibility of a Black-threat association. We begin with an overview of the DIPM, review work suggesting that Black Americans are associated with threat, identify limitations in that work, and present five studies that overcome those limitations.

The Dual Implicit Process Model

The DIPM integrates dual process models of attitudes (e.g., Fazio, 1990; Gawronski & Bodenhausen, 2006) with an evolutionary-derived literature on threat detection (e.g., Blanchette, 2006; LeDoux, 1996, 2012; Öhman, & Mineka, 2001, 2003). The threat literature suggests that organisms that were faster to detect and react to threats to immediate bodily harm were more likely to survive and, consequently, a neural threat-system evolved that advantages the processing of survival threats, both phylogenetic (e.g., snakes, spiders) and ontogenetic (e.g., guns, knives), relative to nonthreatening stimuli. That advantage manifests in faster and stronger perceptual, physiological, and behavioral responses to threats (for details see March et al., 2018a). March et al. (2017), for example, empirically derived images that were threatening (e.g., snarling predators, gunmen), nonthreatening-negative (e.g., injured kitten, feces), positive (e.g., puppy, babies), or neutral (e.g., doorknob, mug) and found that the threatening images were (a) more quickly detected in an embedded image task, (b) more frequent targets of initial eye-gaze, and (c) stronger elicitors of startle-eyeblink. Although dual-process models allow for the implicit processing of valence, they cannot account for the greater processing advantage of threat—a deficit the DIPM rectifies.

The DIPM describes two serially linked implicit processes that can each influence subsequent controlled processing. The first solely attends and initiates responses to immediate survival threats. Such responses involve reactions to promote survival (e.g., reflexive freezing and defensive fighting, autonomic arousal, and amygdala and adrenal activity) and downstream

information processing directed toward the threat. The second, which is well articulated by extant dual-process models, attends to the full evaluative continuum (negative-positive).

As March et al. (2018a) propose, the DIPM's distinction—threatening things are negative, but not all negative things are threatening—affords an important implication for bias. Automatic reactions to social groups could be driven by threat, valence, or both, depending on whether the group is associated with physical threat (e.g., violence, criminality). Encountering a group (or member) associated with threat would activate the threat response and immediate perception and behavior would be geared toward self-preservation. However, the threat response would remain dormant if the group is not associated with violence and, thus, immediate responding could be influenced only by valence processing. Of course, a group associated with both threat and negativity could activate an initial threat response and a subsequent negative valence response both prior to more controlled responses. The important point from the DIPM is that threat is distinguishable from valence with unique downstream consequences geared toward self-preservation. Indeed, when a police officer's dilemma (i.e., immediate shoot behavior) is considered from the DIPM perspective, it is clear that threat processing could yield split-second reactions that differ drastically from that of valence processing. This view is consistent with work suggesting that threat and valence processing are rooted in different learning mechanisms and neural systems and have unique implications for bias (Amodio, 2014, 2019; Amodio & Ratner, 2011). The critical issue for the activation of a self-preserving threat response to social groups is whether or not the social group is associated with threat. The focus of the current work is not on the threat *response*, but on whether White Americans automatically *associate* Black men with survival threat, or simply in terms of negative valence. As we review next, evidence is consistent with the possibility that White Americans implicitly associate Black men with survival

threat.

A Potential Automatic Black-Threat Association

Trait ratings indicate that criminality, hostility, and violence are components of the cultural stereotype of (but not necessarily personal beliefs about) Black Americans (Cottrell et al., 2005; Devine & Elliot, 1995; Krueger, 1996). Acknowledgment of the Black-violence stereotype correlates with the aforementioned laboratory shooter-bias against Black targets (Correll et al., 2002, 2006; Sadler et al., 2012). Priming research suggests that images of Black men facilitate White perceivers' processing of aggressive cues. White Americans, for example, are (a) more likely to construe an ambiguous behavior (i.e., a push) as violent when enacted by a Black than a White man (Duncan, 1976), (b) faster to identify speeded presentations of guns than tools or toys when those objects are preceded by Black than White faces (Kubota & Ito, 2014; Payne, 2001; Todd et al., 2016; Thiem et al., 2019), and (c) require less information to detect degraded images of crime-relevant objects (e.g., gun, knife) than crime-irrelevant objects (e.g., phone, camera) when primed by Black than White faces (Eberhardt et al., 2004). Emotion identification tasks suggest that when deciding whether a face is angry or happy, White Americans are faster to identify angry faces as angry and slower to identify happy faces as happy when the faces are Black than White (Hugenberg, 2005; Hugenberg & Bodenhausen, 2003). Dot-probe tasks suggest that White individuals' visual attention is drawn more to Black than White faces (a) to the extent they associate Blacks more than Whites with danger (Donders et al., 2008), (b) are primed with crime-relevant stimuli (Eberhardt et al., 2004), and (c) the faces display a direct, but not averted, eye gaze (i.e., a sign of threat; Trawalter et al., 2008).

Is it Really Black-Threat?

Two limitations undermine confidence in the conclusion that White Americans evaluate

Black Americans with survival threat. Because threatening stimuli are negative but negative stimuli are not necessarily threatening (i.e., "negative" is the broader umbrella category), operationalizing threat without also operationalizing negativity prevents empirical distinction of threat and valence processing. For example, the aforementioned tendency for guns to be identified faster and with less information than are tools and toys when primed by Black than White faces could be due to a Black-negative instead of a Black-threat association because the studies did not concurrently examine race primes on nonthreatening-negative objects. Indeed, White participants are slower to identify positive words as "good" and faster to identify negative words as "bad" when primed by Black than White faces (Dovidio et al., 1997; Fazio et al., 1995). Similarly, the correlation between the Black-violent stereotype and speeded shooter-bias could be a product of a Black-negative association because the studies did not concurrently assess nonthreatening-negative components of the Black stereotype (e.g., "lazy"). Without operationalizing both threat and negativity, most existing studies do not distinguish a Black-threat and Black-negative association (for further discussion see March et al., 2020).

Three exceptions are Donders et al. (2008), Hugenberg (2005), and Judd et al. (2004; also see replication by Todd et al., 2016). Consistent with a Black-threat association, Donders et al. (2008) demonstrated that among White participants, a Black-danger stereotype but neither a Black-danger-irrelevant stereotype nor Black-negative association predicted greater attentional capture of Black than White faces. Hugenberg (2005), in contrast, suggests that the tendency to more quickly identify anger on angry Black than angry White faces is driven by a Black-negative rather than Black-threat association because the same speeded tendency occurs with the identification of sadness on sad Black than sad White faces. Of course, this does not exclude the possibility of both Black-threat and Black-negative associations; it could have proven useful to

also have trials in which threat and negative valence were pitted against each other by having participants determine whether angry and sad faces, respectively, are angry or sad (rather than angry vs. happy, or sad vs. happy). Finally, Judd et al. (2004) demonstrated that Black (relative to White) faces facilitate the identification of guns, but not the identification of insects, which is consistent with a Black-threat but not a Black-negative association. Judd et al., however, suggest the effect is due to a semantic association of Black and Gun (i.e., stereotype) rather than an evaluative association given that Black (relative to White) faces also facilitated the identification of sports objects (i.e., a positive component of the stereotype of Blacks), but not fruit. What is not clear, however, is what would have occurred had their task required participants to evaluate the targets (i.e., guns, insects, sports equipment, fruit) as "Good" or "Bad," rather than semantically label them as "Gun" or "Insect" and "Sports" or "Fruit."

The second limitation is the use of a single outgroup. If White participants differentially react to Black than White targets, it is unclear whether the reaction is unique to Black targets or an intergroup reaction that occurs across outgroups. For example, regarding the tendency toward greater attentional capture of Black than White faces, White participants similarly evidence greater attentional capture of Asian than White faces (Al-Janabi et al., 2012). Two exceptions are Cottrell et al. (2005), who indicate that White participants report more fear and threat to physical safety from Black Americans than from other groups, and Amodio et al. (2003), who suggest that the startle eyeblink is facilitated by Black but not Asian faces.

So, the literature provides data consistent with the possibility that Black Americans are implicitly associated with a survival threat. The data, nonetheless, are also consistent with the possibilities that the association is with negativity (not threat per se) and reflects a more general intergroup process that occurs across outgroups rather than being unique to White American's

perception of Black men.

Current Research

We present five studies that collectively overcome those limitations to test if White Americans uniquely and automatically associate Black Americans with threat, negativity, or both, and the latter three studies include trials that directly pit threat against negativity to assess which association is stronger. Studies 1 and 2 assess latency to evaluate threatening, negative, and positive target images as good versus bad when primed by Black versus White male-faces. To the extent that Whites implicitly associate Black men with threat (and/or negativity) participants should be faster to evaluate threatening (and/or negative) targets as bad when primed with Black than White faces. Studies 3 and 4 use mouse-tracking to assess how early in the decision process White Americans begin deciding whether Black and White (and, in Study 3, Asian) male-faces displaying anger, sadness, happiness, or no emotion are, in Study 3, dangerous, depressed, cheerful, or calm, or in Study 4, dangerous, negative, or positive. To the extent Whites associate Black men with threat (and/or negativity), they should be biased earlier in the decision process toward categorizing Black faces as dangerous (and/or depressed/negative) relative to White (and Asian) faces. Trials in which dangerous and depressed/negative are paired together directly test the relative strength of the threat vs. negative association . In contrast to Studies 1-4 that utilize images as primes and/or targets, Study 5 assesses latency to evaluate threatening and negative words as “dangerous” versus “negative” when primed by Black versus White names. To the extent that Whites associate Black men with threat (and/or negativity) participants should be faster to evaluate threatening (and/or negative) words as dangerous (and/or negative) when primed with Black than White faces.

Before transitioning to the studies, we emphasize two points. First, we use latencies in evaluative priming and mouse-tracking tasks to assess a possible Black-threat association differentiated from a Black-negative association. Those associations should not be confused with the threat-response delineated by the DIPM. The threat response involves perceptual, physiological/neural, and behavioral reactions geared toward self-preservation. Threatening stimuli should elicit earlier and stronger responses than nonthreatening stimuli to the extent to which the particular responses are in the service of self-protection. Latencies to button-presses and mouse-movements in the evaluative priming and mouse-tracking tasks are not self-protective reactions. Those latencies assess the relative strength of association of Black vs. White with threat, negativity, and positivity, with earlier latencies indicating a stronger association with one race than the other. A threat-association is necessary for a group to elicit a threat response. The purpose of the current work is to assess whether White Americans associate Black-men with survival threat.

Second, all of our participants are White Americans because our hypothesis concerns White Americans' perceptions of Black men. This is not to suggest that White Americans are the only persons to associate Black men with survival threat. As March et al. (2018) discuss, the DIPM raises the possibility that members of groups stereotyped as violent might themselves process their ingroup in terms of threat and positivity (with the latter via ingroup favoritism mechanisms). Consequently, Black-Americans might similarly manifest a Black-threat association. Testing that, however, is complicated by the possibility that they might also associate White-Americans with survival threat due to the history of Whites' violence toward Blacks in the US. Consequently, we confine participation to White Americans.

Studies 1 and 2

Evaluative priming tasks have been used to implicitly assess whether participants differentially associate Blacks and Whites with negativity and positivity by comparing the extent to which Black vs. White primes influence the speed with which negative and positive targets are evaluated as good or bad (e.g., Dovidio et al., 1997; Fazio et al., 1995). We modified the task by methodologically distinguishing among threatening, negative and positive targets with a 2(Prime: Black, White face) x 3(Target: threatening, negative, positive) within-subjects design. Importantly, the response label "bad" is applicable to both threatening and negative targets and differential response times to those targets as a function of the race prime enables us to assess whether White participants associate Black (vs. White faces) with threat, negativity, or both.

Study 2 is a direct replication of Study 1 and we present them in parallel. We determined sample size for Study 1 by the number of participants obtained by the end of the semester. We determined sample size for Study 2 via a power simulation (Soderberg et al., 2018) of the Study 1 data. The simulation indicated that 120 participants would provide 90% power to detect the Black-versus-White threat effect and 99% power to detect the 2(Prime) x 3(Target) interaction (we oversampled by 10% to allow for unusable data).

Methods

White undergraduates (Study 1: N = 81, 51 females, 1 unspecified; Study 2: N = 132, 89 females) participated for partial credit in an introductory psychology course and sat in separate cubicles with a 48cm high-speed, high-resolution monitor and computer. Instructions explained that pairs of pictures would be presented sequentially with the first being a face and the second the target, and they should indicate as quickly and accurately as possible whether the target is bad or good by pressing the "Z" or "/" key, respectively. They practiced eight trials to adapt to

the task and transitioned with a button click to complete 256 trials before being debriefed. Each trial began with a centrally located mosaic image for 500ms that functioned as a fixation and pre-mask, which was replaced for 200ms by a face, which was replaced for 200ms by a target, which was replaced for 100ms by the mosaic post-mask and ended on response to the prompt of whether the target was bad or good. A 1500ms blank screen separated trials.

As primes, we used 30 Black and 30 White male neutral-faces from the Chicago Face Database (Ma et al., 2015) after cropping each to 500x500-pixels confined vertically between the eyebrows and bottom lip and horizontally between the outsides of the eyes. As targets, we used March et al.'s (2017) images of threat, nonthreatening-negativity, and positivity (30 of each; see Figure 1 for example stimuli and supplemental materials for all stimuli), which (as described by March et al.) were validated through extensive pilot testing such that the threatening and negative sets were both rated (on 1 to 7 scales) as low in positivity ($M_{\text{threat images}} = 1.99$, $M_{\text{negative images}} = 1.48$), high in negativity ($M_{\text{threat images}} = 4.45$, $M_{\text{negative images}} = 4.95$), and differed in threat ($M_{\text{threat images}} = 5.78$, $M_{\text{negative images}} = 3.20$). The negative set contains a broad swath of nonthreatening negative images which, in line with earlier work in this domain (e.g., Donders et al., 2008; Judd et al., 2004; Todd et al., 2016), captures a negative, but nonthreatening, evaluation (i.e., objects that evoke antipathy, dislike; Rozin, 1986). Additionally, threat and negative stimuli sets were both rated (on 1 to 7 scales) high in arousal ($M_{\text{threat images}} = 5.85$, $M_{\text{negative images}} = 6.12$).

To prevent a response bias of "bad," 50% of trials displayed a positive target, 25% displayed a threatening target, and 25% displayed negative targets (i.e., 50% of trials required a response of "good" and 50% "bad"; practice trials had the same structure). Each target type was primed by equal proportions of Black and White faces. The order and pairing of prime and target

were randomized with all targets presented once before any was re-presented.

Figure 1. Example of target stimuli used in Studies 1 and 2.

Response latency to a correct response (i.e., bad for threatening/negative targets, good for positive targets) was the dependent measure. Four targets in Study 1 (1 positive, 3 negative) and one in Study 2 (negative) had error rates above 20% (our a priori cutoff). We excluded responses to those error prone targets (Study 1: n = 875, 4.2%; Study 2: n = 275, 0.8%) and all remaining incorrect responses (Study 1: n = 415, 2.1%; Study 2: n = 590, 1.7%) leaving 19,446 and 32,927 correct responses in Study 1 and 2. We subsequently excluded slow times exceeding three interquartile ranges of the 75th percentile (Tukey, 1977; Study 1: n = 610, 3.1%; Study 2: n = 809, 2.4%), and two participants (Study 1) with less than 70% of their data remaining, yielding

79 participants (18,514 responses) in Study 1 and 132 participants (32,118 responses) in Study 2.

Results

A 2(Prime: Black, White) x 3(Target: Negative, Positive, Threatening) multivariate repeated measures ANOVA on natural logged times revealed an interaction in Study 1, $F(2, 77) = 17.29$, $p < .0001$, $f = .67$ and Study 2, $F(2, 130) = 16.86$, $p < .0001$, $f = .51$. As displayed in Figure 2, Black (versus White) faces yielded (a) faster response to threatening targets [Study 1: $F(1, 78) = 7.14$, $p = .009$, $d_z = -0.30$; Study 2: $F(1, 131) = 5.34$, $p = .0225$, $d_z = -0.20$], (b) slower response to positive targets [Study 1: $F(1, 78) = 40.69$, $p < .0001$, $d_z = 0.72$; Study 2: $F(1, 131) = 34.73$, $p < .0001$, $d_z = 0.51$], and (c) no difference to negative targets [Study 1: $F(1, 78) = 0.36$, $p = .551$, $d_z = -0.07$; Study 2: $F(1,131) = 0.30$, $p = .5850$, $d_z = -0.05$].[1]

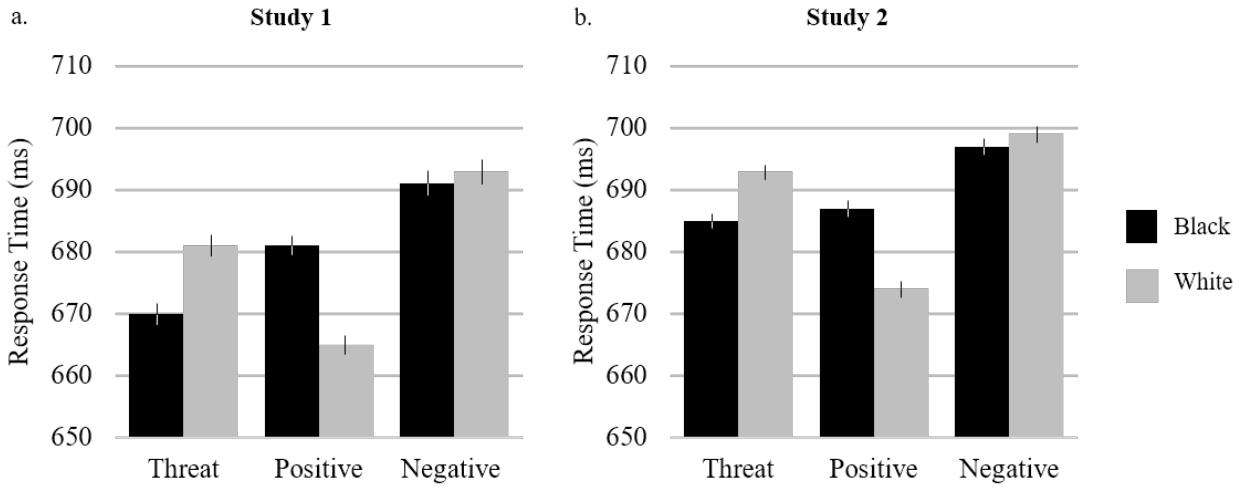


Figure 2. Mean response time as a function of Prime and Target for (a) Study 1 and (b) Study 2. Error bars are ±1 *SEM* calculated within-participants (O'Brien & Cousineau, 2014).

Discussion

By methodologically differentiating threat, negativity, and positivity, Studies 1 and 2 were able to distinguish evaluative race-associations in terms of threat and valence. Participants were faster to evaluate threatening targets, but not negative targets, when primed by Black than

White faces. Importantly, the facilitating effect of Black vs. White faces on threatening targets did not differ as a function of whether those targets were guns versus threatening animals, [Study 1: $F(1, 78) = 0.01$, $p = .9397$, $f = .009$; Study 2: $F(1, 131) = 0.82$, $p = .3673$, $f = .079$]. This indicates that the Black (relative to White) threat-association is not merely a stereotype about weapons or a semantic association between Black and gun (e.g., Judd et al., 2004; Todd et al., 2016), and is more broadly an association between Black and threat. These data suggest that threat, but not general negative valence, is a primary source of automatic anti-Black bias (e.g., Donders et al., 2008). Indeed, following Black primes participants were faster to identify threatening than negative targets, [Study1: $F(1, 78) = 23.35$, $p = .0001$, $d_z = -0.54$; Study 2: $F(1, 131) = 21.88$, $p = .0001$, $d_z = -0.41$]. Readers interested in whether the association of Black vs. White-threat is stronger than that of Black vs. White-negative should note that when we collapse across studies, the significant effect of Black vs. White primes on threatening targets is stronger than the non-significant effect of Black vs. White primes on negative targets, $F(1, 209) = 3.80$, $p = .0525$, $f = .135$ (i.e., Black vs. White x Threat vs. Negative)^2, and does not vary by study, $F(1, 209) = 0.11$, $p = .7380$, $f = .023$ (i.e., Black vs. White x Threat vs. Negative x Study 1 vs. 2). Although agnostic when designing Study 1 (and 2) as to whether Black-threat would be stronger than Black-negative, we include in Studies 3, 4, and 5 a procedure that provides a direct test of their relative influence.

Two limitations are of note. The lack of a second outgroup maintains the possibility that the data reflect general outgroup associations, not associations unique to Black men. Also, 12 of the 30 threat targets were gunman, of whom 8 were White and 4 Black. Notably, response time as a function of the Black vs. White face primes did not differ for Black vs. White gunman, [Study 1: $F(1, 78) = 0.67$, $p = .4169$, $f = .092$ Study 2: $F(1, 131) = 0.90$, $p = .3443$, $f = .083$

Nonetheless, it would have been ideal had gun images not displayed race.[3] To address these limitations, we conducted Studies 3 and 4 with a different paradigm (mouse-tracking) that did not utilize gunman (nor the other target) images and included two outgroups, and Study 5 (evaluative priming) without any image stimuli.

Study 3

Study 3 uses mouse-tracking to test the association of Black-males (vs. White and Asian males) with threat, negativity, and positivity, and directly tests whether Black-threat is stronger than Black-negative via trials that pit threat against negativity. Mouse-tracking records XY coordinates of motion when participants move the mouse to categorize a stimulus in regard to response alternatives (i.e., a target label and a distractor label) and provides information, beyond total response time, about the influence of competing alternatives during the decision process (Freeman, 2018; Hehman et al., 2015). That is, the response path reflects the association strength of the stimulus with the target-label and the distractor-label. White participants viewed White, Black, and Asian male-faces that were angry, sad, happy, or emotionally neutral, and chose one of two labels to describe the face: a target label that accurately described the face (i.e., Dangerous for angry, Depressed for sad, Cheerful for happy, and Calm for neutral) and a distractor label that did not accurately describe the face (i.e., one of the remaining labels, e.g., Depressed, Cheerful, or Calm for angry).

If White Americans uniquely associate Blacks with threat (or negativity), they should be begin categorizing Black faces as dangerous (or depressed) earlier in the decision process than White and Asian faces. In particular, for angry (or sad) faces, participants should be less affected by the distractor label and begin moving earlier in time to the target-label dangerous (or depressed) if the face is Black than White or Asian. For faces that are not angry (or sad)

participants should be more affected by the distractor-label dangerous (or depressed) and begin moving later in time to the target label if the face is Black than White or Asian. Furthermore, unlike prior research on race and facial emotion (e.g., Hugenberg, 2005) and the evaluative priming paradigm of Studies 1 and 2, the current study involves trials that directly pit threat against negativity to assess their relative influence. If White Americans associate Blacks more strongly with threat than negativity (as Studies 1 and 2 suggest), categorization of a sad Black (vs. White or Asian) face as Depressed should be delayed when Dangerous is a response option. Alternatively, if White Americans associate Blacks more strongly with negativity than threat, categorization of an angry Black (vs. White or Asian) face as Dangerous should be delayed when Depressed is a response option. We determined sample size by the number of participants we could run in a semester.[4]

Methods

White undergraduates (N = 118, 85 females, 1 unspecified) participated for partial credit in an introductory psychology course and sat in separate cubicles with a 48cm high-speed, high-resolution monitor and computer. Instructions explained that future studies require face pictures that can be quickly and accurately identified as calm, cheerful, dangerous, or depressed. Those labels were defined such that calm faces "look emotionless, neutral, flat," cheerful faces "look happy, friendly, joyful," dangerous faces "look angry, scary, threatening," and depressed faces "look sad, gloomy, unhappy." Participants were told that each trial would display a start-button on the bottom-center of the screen and two expression-labels at the upper left and right corners, respectively. Upon clicking "start," a face would appear above the button, and they would move the mouse to click the label that describes the face. Participants were reminded of the need for quick and accurate identification and had to begin moving the mouse when they clicked start

(rather than deciding and then moving to a label), with warnings noting if they took too long to begin moving (> 300ms) or click a label (> 2000ms; see Figure 3). Participants practiced 10 trials in which they categorized pictures of food as “fruit” or “vegetable,” and then completed six blocks (60 trials each) of categorizing angry, happy, neutral, and sad faces before being debriefed.

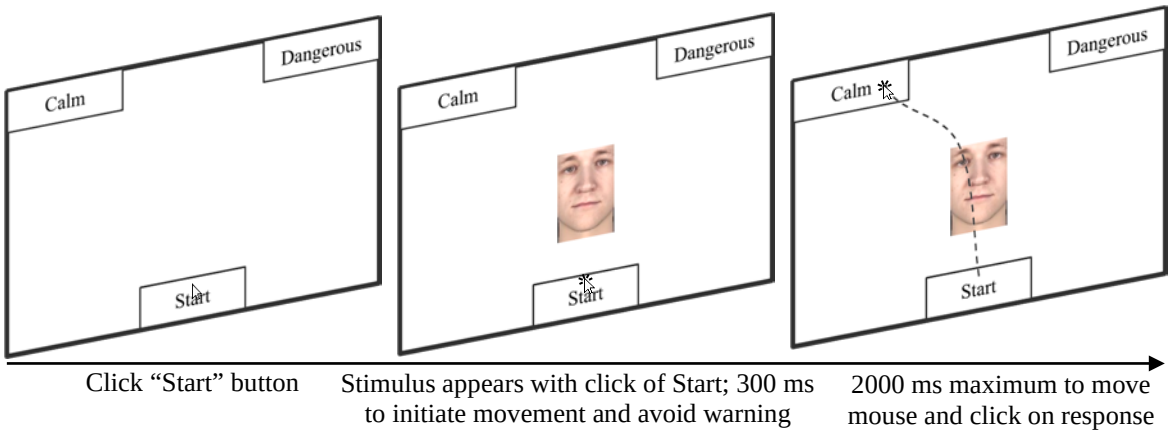


Figure 3. Trial time course.

Each block presented two labels (e.g., dangerous and depressed) and facial expressions relevant to those labels (e.g., angry and sad). The paired labels (and expressions) changed across blocks. On any trial, one label (target) described the expression and the other (distractor) did not. Within a block an equal number of Black, White, and Asian faces displayed each expression, and halfway through a block the left-right position of the labels changed (with a preceding screen noting the change). Across the 360 trials, each label was paired with each of the other labels as a target and distractor, and there were 10 trials of each target and expression paired with each distractor for each race (e.g., 10 neutral White faces with Calm and Happy).[5] The order of blocks, faces within blocks, and left-right starting position of labels was randomized.

To create the expressions, we began with 60 male neutral-faces (20 of each race) from the Chicago Face Database (Ma et al., 2015) and imported each into a morphing program (FaceGen)

that uses the Facial Action Coding System (Ekman & Friesen, 1978) to produce digitized angry, happy, neutral, and sad versions of each imported face (see supplemental materials). This yielded 240 faces which were cropped to 450x650 pixels. We pilot tested the faces (see supplemental materials) to ensure that the expressions were perceived as intended, which yielded 10 models of each race whose four expressions we used in the main study.

Data were recorded in MouseTracker (www.mousetracker.org), which mapped each trial on a standard x, y coordinate space in 20ms intervals equating the left/right position of the target label, and exported to SAS for further processing. Of the 41,037 trials (see footnote 4), 588 (1.43%) did not have recorded data because participants exceeded the 2000ms limit and 761 (1.85%) ended incorrectly (i.e., participants clicked the distractor, not target, label). Of the 39,688 useable trials, based on our a priori criteria, we excluded 2,487 (6.27%) in which movement initiated too late (>300ms of clicking start) and 1,407 (3.54%) that ended abnormally fast (<600ms), yielding 35,794 trials from 118 participants.

Our hypothesis relevant interest is the participant's time of initiating correct categorization (i.e., TICC; March & Gaertner, 2021) of the emotive face in regard to the target (vs. distractor) label, and particularly whether that time of initiating categorizing varied by race. For example, what was the time of initiating correct categorization of the angry face as Dangerous (i.e., target label) relative to Sad, Cheerful, or Calm (i.e., distractor labels), and was it earlier when the angry face was Black than White or Asian? We provide an overview of TICC and the interested reader should consult March and Gaertner (2021) for more details (including software code). For stable mouse-trajectory estimates, we averaged the time-synched x-, y-coordinates across the (up to) 10 trials of each target/distractor pairing of a race (e.g., 10 trials of sad Black-faces with Depressed and Happy) for each participant. This yielded 36 average

trajectories (12 of each race) for each participant (of the possible 4,248 trajectories from 118 participants, 22 were missing from 7 participants due to trials exceeding 2000ms). For each average trajectory, we calculated the Euclidean distance of the mouse at each time interval from the target and distractor labels, respectively.[6] Euclidean distance from each label is necessary because pure vertical movement brings the mouse equally closer to both labels, and horizontal distance is insufficient in that horizontal movement lower on the screen is further from a label than is the same movement higher on the screen. Thus, we calculated at each time interval the difference in the Euclidean distance from the target and distractor label, which yields a sigmoid curve over time as depicted in Figure 4.

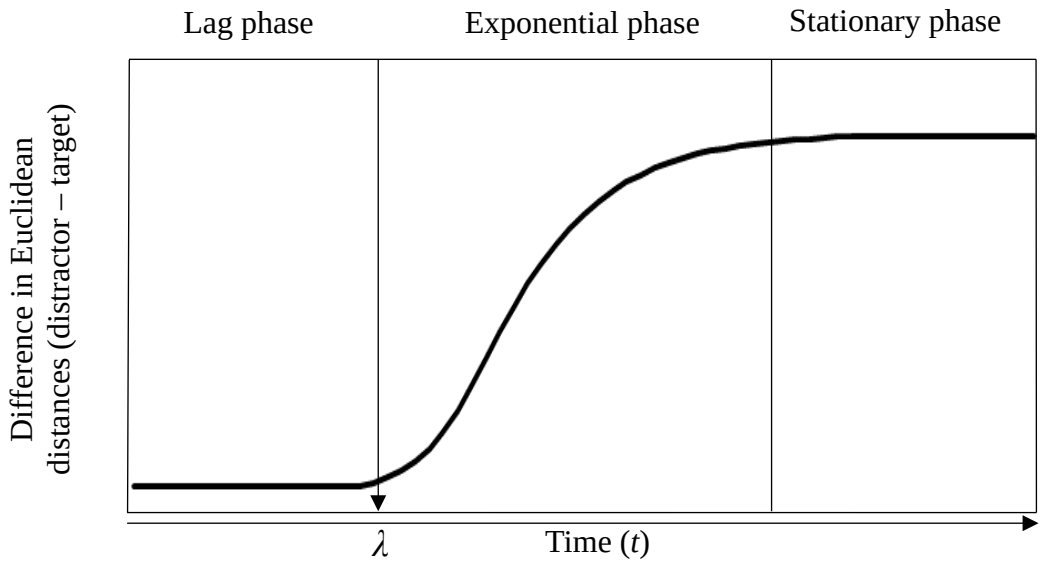


Figure 4. Sigmoid for difference in Euclidean distances over time.

The flat part of the curve early in time is vertical movement bringing the mouse equally closer to the target and distractor. The exponential slope is movement closer to the target and further from the distractor, which asymptotes later as the mouse reaches the target. A sigmoid over time occurs for many phenomena, such as bacterial growth. The phases on top of Figure 4 are what bacterial scientists refer to as the lag phase in which growth is dormant, the exponential

phase in which growth multiplies, and the stationary phase in which growth is maximized. The lambda symbol (λ) on the time-axis is, for bacterial scientists, the point in time when bacteria transition from dormancy to exponential growth. For us, λ is the point in time when participants begin moving relatively closer to the target than distractor label (i.e., time of initiating correct categorization, or TICC; March & Gaertner, 2021).

Two non-linear models each estimate λ (i.e., TICC; and the other parameters of the sigmoid), i.e., the Gompertz and Baranyi models (Baty & Delignette-Muller, 2004):

$$y_t = y_{min} + y_{max} e\left\{-e\left[\frac{\mu_m e}{y_{max}}(\lambda - t) + 1\right]\right\} \qquad \text{Gompertz}$$

$$y_t = y_{max} + \ln\left(\frac{-1 + e^{\mu_m \lambda} + e^{\mu_m t}}{(-1 + e^{\mu_m t}) + e^{(\mu_m \lambda + y_{max} - y_{min})}}\right) \qquad \text{Baranyi}$$

In each model, y_t is the Euclidean-distance difference at a given time (t), y_{min} is the lower asymptote, y_{max} is the upper asymptote, μ_m is the maximum growth rate, e is a mathematical constant ≈ 2.718 (i.e., Euler's number), and λ is the TICC. Software capable of nonlinear regression can estimate the parameters of the Gompertz and Baranyi models from each participant's Euclidean-distance difference at each time point. We used SAS Proc NLIN to fit the Gompertz and Baranyi models to the time by Euclidean-distance difference (i.e., sigmoid) curve for each of the 12 target/distractor pairings of each race for each participant. Both models converged on 4,199 of the 4,226 curves (99.36%) and evidenced exceptional fit with an average pseudo-R^2 = .9428. With no reason to prefer the Gompertz vs. Baranyi model, we averaged their TICC estimate for each of the 12 target-distractor pairings for each race for each participant.

Results

To test whether White Americans are biased earlier in the decision process for Black (than White or Asian) faces by threat and/or negative valence, we entered participants' TICC estimate for a given target-label into a 3 (Distractor) x 3 (Race) multivariate repeated-measures

ANOVA (degrees of freedom vary due to missing TICC estimates from non-convergence or missing trajectories as described above). Figures 5-8 display the average mouse-trajectory and Euclidean-distance differences over time (with area-of-focus on TICC) of each race for each target/distractor pairing.[7]

Dangerous-target (angry face). TICC for an angry face was influenced by a race main effect $F(2, 107) = 19.82$, $p < .0001$, $f = .61$, that was not moderated by the distractor type (i.e., Race x Distractor), $F(4, 105) = 1.09$, $p = .3632$, $f = .20$ (Figure 5). Regardless of the distractor, participants began moving *earlier* in time to Dangerous when the angry face was Black (M = 499ms) than White (M = 524ms), $F(1, 108) = 16.82$, $p < .0001$, $d_z = -0.41$, or Asian (M = 538ms), $F(1, 108) = 38.70$, $p < .0001$, $d_z = -0.58$, and they began moving earlier to the White than Asian face, $F(1, 108) = 4.64$, $p = .0334$, $d_z = -0.22$. Stated in regard to threat and valence, the tendency to begin moving earlier in time to the threatening target (Dangerous) for an angry Black than White or Asian face was no more affected by the negative distractor (Depressed) than the positive (Happy) or neutral (Calm) distractors. That is, participants began categorizing angry Black faces as Dangerous earlier than angry White or Asian faces regardless of the distractor label.

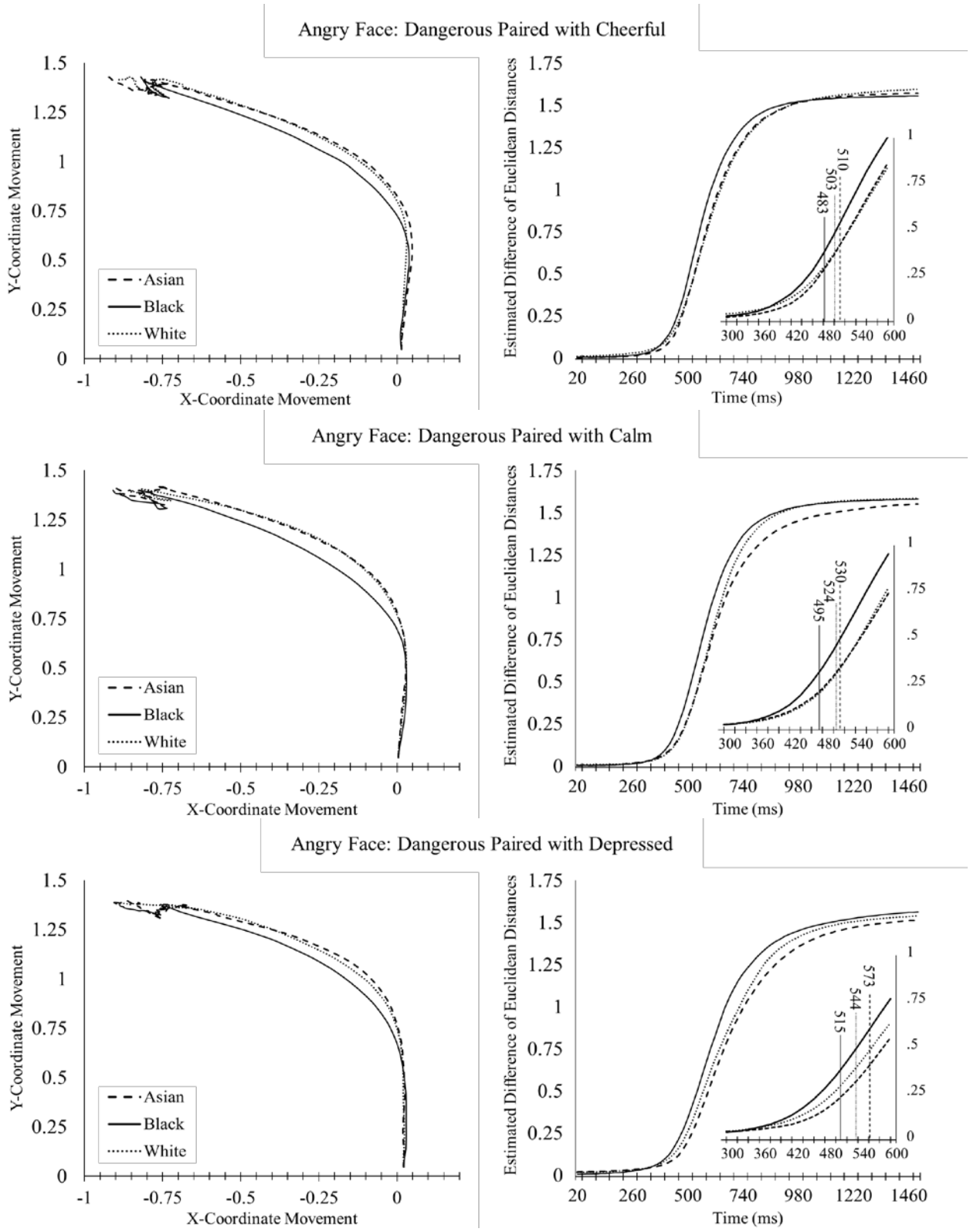


Figure 5. Mean mouse-trajectory (left panels) and Euclidean-distance differences over time (with area-of-focus on TICC; right panels) to Dangerous target for angry faces as a function of race and distractor.

Depressed-target (sad face). TICC for a sad face was influenced by a Race x Distractor interaction, $F(4, 105) = 7.53$, $p < .0001$, $f = .54$ (Figure 6). With the Dangerous distractor, participants began moving *later* in time to Depressed when the sad face was Black (M = 576ms) than White (M = 521ms), $F(1, 108) = 15.73$, $p = .0001$, $d = 0.35_z$, or Asian (M = 522ms), $F(1, 108) = 10.68$, $p = .0015$, $d_z = 0.32$, and the latter two did not differ $F(1, 108) = 0.00$, $p = .9714$, $d_z = -0.02$. With the Cheerful distractor, participants began moving earlier in time to Depressed when the sad face was Black (M = 467) than White (M = 512ms), $F(1, 108) = 10.76$, $p = .0014$, $d_z = 0.33$, or Asian (M = 498ms), $F(1, 108) = 7.68$, $p = .0066$, $d_z = 0.27$, and the latter two did not differ, $F(1, 108) = 1.13$, $p = .2911$, $d_z = 0.11$. With the Calm distractor, time of movement to Depressed did not differ among the races, Fs$(1, 108) < 1.19$, ps $> .278$, d_zs < 0.14. Stated in regard to threat and valence, the threatening distractor (Dangerous) interfered with the time participants began categorizing the sad face with the negative label (Depressed) more for Black than White or Asian faces and participants began categorizing sad Black faces later in time than sad White or Asian faces.

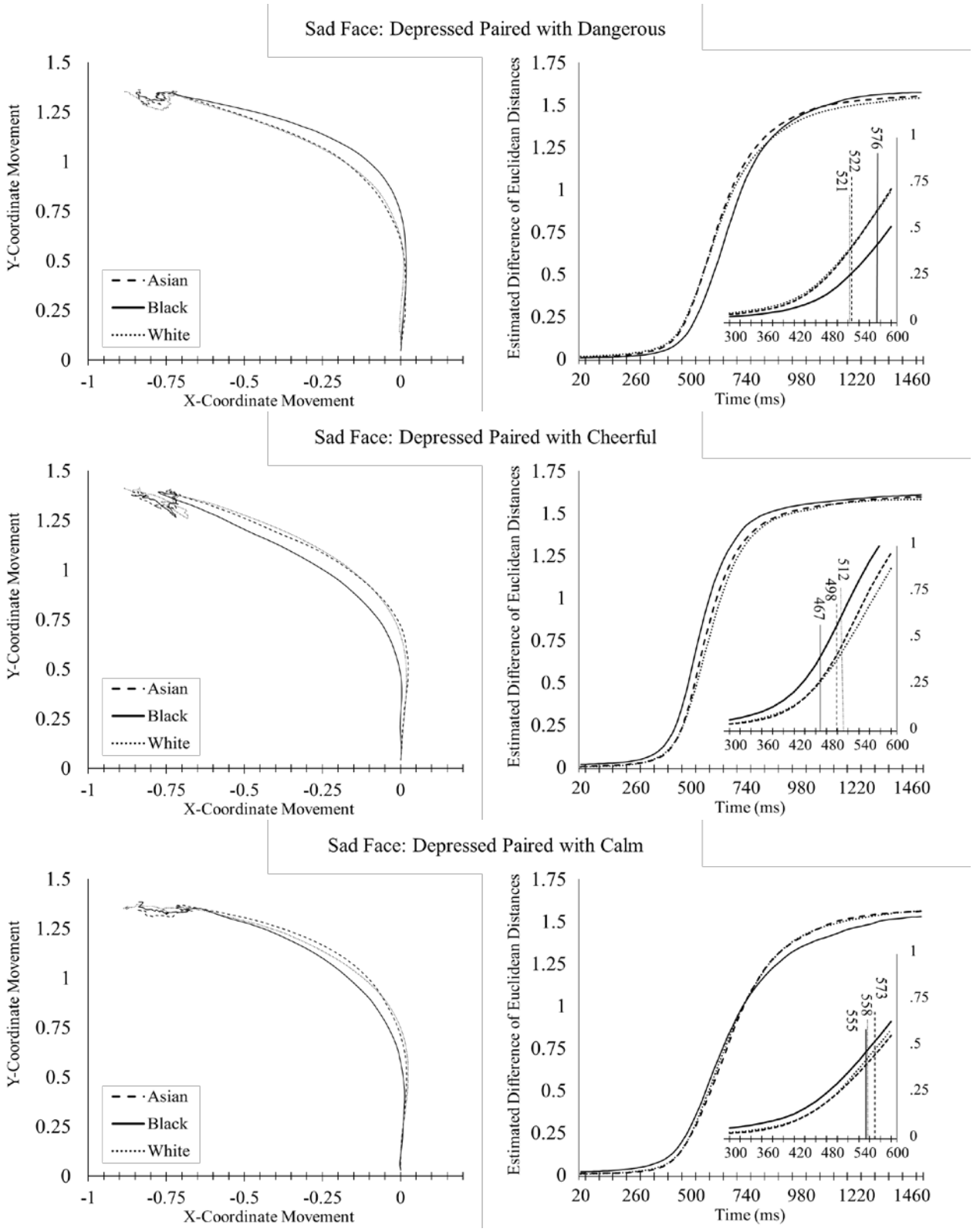


Figure 6 Mean mouse-trajectory (left panels) and Euclidean-distance differences over time (with area-of-focus on TICC; right panels) to Depressed target for sad faces as a function of race and distractor.

Cheerful-target (happy face). TICC for a happy face was influenced by a Race x Distractor interaction, $F(4, 106) = 4.70$, $p = .0016$, $f = .42$ Figure 7). With the Dangerous distractor, participants began moving later in time to Cheerful when the happy face was Black (M = 548ms) than White (M = 481ms), $F(1, 109) = 50.34$, $p < .0001$, $d_z = 0.57$, or Asian (M = 492ms), $F(1, 109) = 23.52$, $p < .0001$, $d_z = 0.42$, and the latter two did not differ, $F(1, 109) = 1.29$, $p = .2577$, $d_z = 0.06$. Similarly, with the Depressed distractor, participants began moving later in time to Cheerful when the happy face was Black (M = 548ms) than White (M = 499.73ms), $F(1, 109) = 11.91$, $p = .0008$, $d_z = 0.33$,or Asian (M = 499.68ms), $F(1, 109) = 11.71$, $p = .0009$, $d_z = 0.33$, and the latter two did not differ, $F(1, 109) = 0.00$, $p = .9969$, $d_z = 0.02$. With the calm distractor, time of movement to Happy did not differ among the races, Fs(1, 109) < 2.30, ps > .132, d_zs < 0.16. Stated in regard to threat and valence, both the threatening (Dangerous) and negative (Sad) distractors interfered with the time participants began categorizing the happy face with the positive label (Cheerful) more for Black than White or Asian faces.

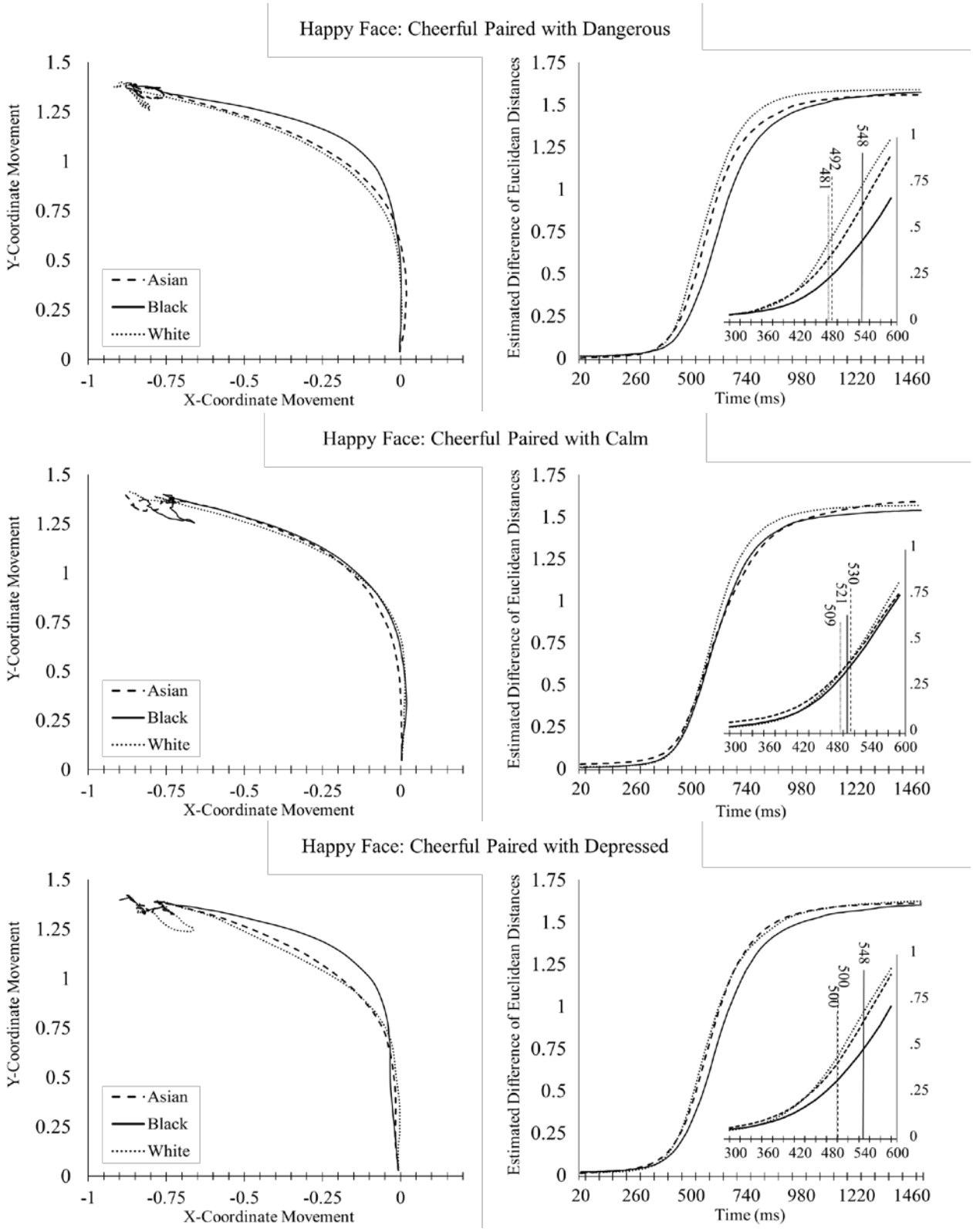


Figure 7. Mean mouse-trajectory (left panels) and Euclidean-distance differences over time (with area-of-focus on TICC; right panels) to Cheerful target for happy faces as a function of race and distractor.

Calm-target (neutral face). TICC for a neutral face was influenced by a Race x Distractor interaction, $F(4, 106) = 4.87$, $p = .0012$, $f = .43$ (Figure 8). With the Dangerous distractor, participants began moving later in time to Calm when the neutral face was Black (M = 567ms) than White (M = 518ms), $F(1, 109) = 21.09$, $p < .0001$, $d_z = 0.46$, or Asian (M = 523ms), $F(1, 109) = 15.93$, $p = .0001$, $d_z = 0.39$, and the latter two did not differ, $F(1, 109) = 0.20$, $p = .6568$, $d_z = 0.07$. With the Depressed distractor, participants began moving later in time to Calm when the neutral face was Black (M = 573ms) than White (M = 529ms), $F(1, 109) = 15.71$, $p < .0001$, $d_z = 0.37$, or Asian (M = 551ms), $F(1, 109) = 4.26$, $p = .0414$, $d_z = 0.20$, and later when the neutral face was Asian than White, $F(1, 109) = 3.68$, $p = .0577$, $d_z = 0.16$. With the Cheerful distractor, time of movement to Calm did not differ among the races, Fs$(1, 109) < 2.05$, ps $> .155$, ds < 0.16. Stated in regard to threat and valence, both the threatening (Dangerous) and negative (Sad) distractors interfered with the time participants began categorizing the neutral face with the neutral label (Calm) more for Black than White or Asian faces.

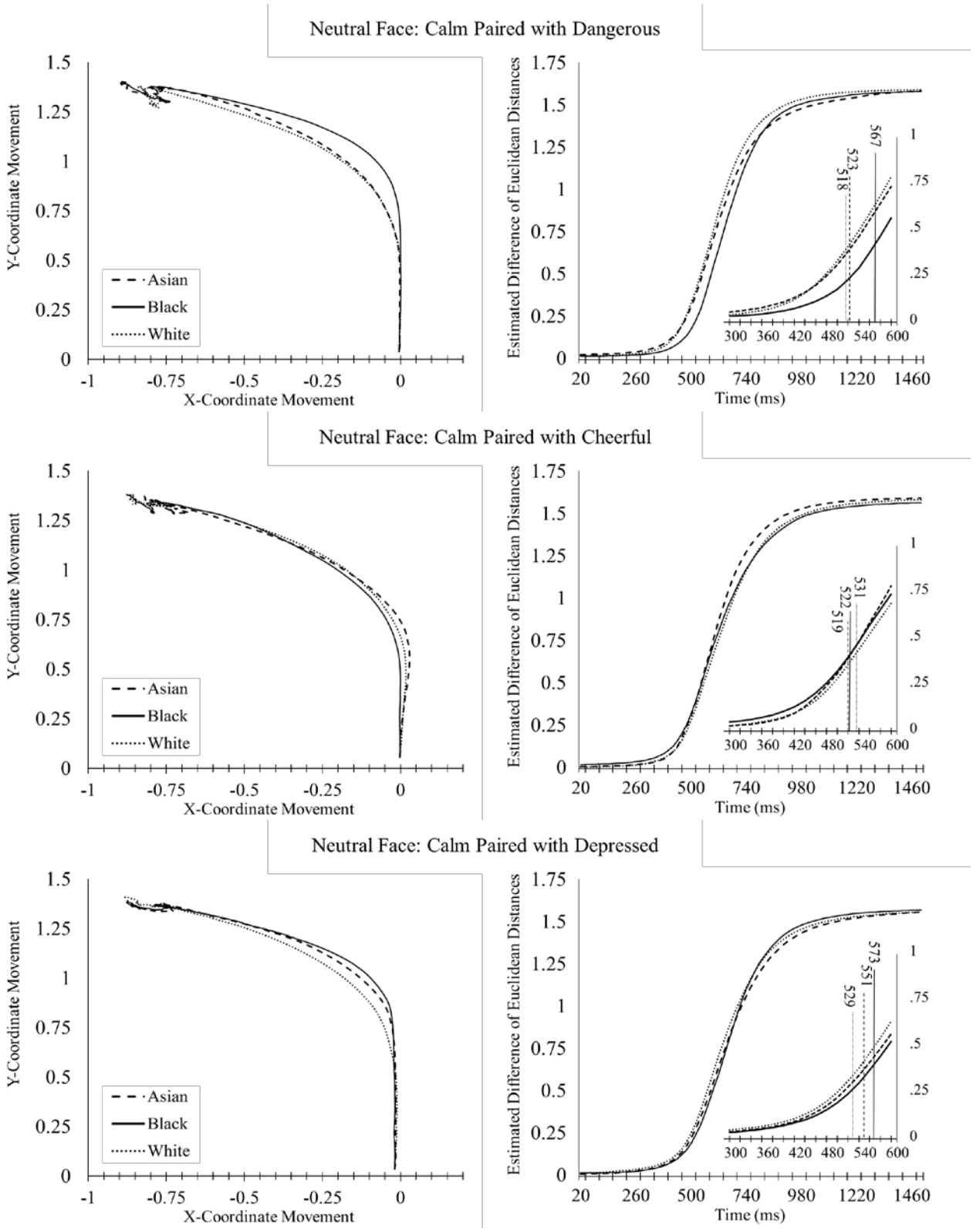


Figure 8. Mean mouse-trajectory (left panels) and Euclidean-distance differences over time (with area-of-focus on TICC; right panels) to Calm target for neutral faces as a function of race and distractor.

Discussion

TICC for neutral and happy faces tell a less nuanced story about threat and valence associations than does TICC for angry and sad faces. The neutral and happy faces suggest that both threat and negativity are more strongly associated with Black males than with White or Asian males. Categorization of neutral and happy faces as Calm and Cheerful, respectively, began later in time for Black than White and Asian faces when Dangerous or Depressed were category options. That is, the threatening (Dangerous) and negative (Depressed) distractors produced more interference for processing neutral and happy Black faces than neutral and happy White and Asian faces.

The angry and sad faces, which pit threat and negative-valence in direct competition via the paired presentation of Dangerous and Depressed as targets and distractors, indicate that the threat association is stronger than the negative-valence association for the perception of Black (than White and Asian) faces. Categorization of angry faces as Dangerous began *earlier* in time for Black than White and Asian faces (regardless of whether Depressed, Cheerful, or Calm were category options, i.e., distractors). However, categorization of sad faces as Depressed began *later* in time for Black than White and Asian faces when Dangerous was a category option (but not when Cheerful or Calm were options). That is, Dangerous interfered as a distractor in the time course of categorizing sad Black (versus White and Asian) faces, but Depressed did not interfere as a distractor in the time course of categorizing angry Black (versus White and Asian) faces. Indeed, analysis of trials involving only angry and sad faces with Dangerous and Depressed labels conceptually replicate the pooled Study 1 and 2 partial-interaction of Black vs. White x Threat vs. Negative, $F(1, 209) = 3.80$, $p = .0525$, $f = .135$, with significant partial-interactions of Black vs. White x Angry-to-Dangerous vs. Sad-to-Depressed, $F(1, 113) = 21.53$, $p = .0001$, f

= .436, and Black vs. Asian x Angry-to-Dangerous vs. Sad-to-Depressed, $F(1, 113) = 23.84$, $p = .0001$, $f = .459$ and the overall 3(Race: Black, White, Asian) x 2(Face: angry-to-Dangerous vs. sad-to-Depressed) interaction, $F(2, 112) = 14.28.$, $p = .0001$, $f = .505$—there is not an analogous effect of White vs. Asian x Angry vs. Depressed, $F(1, 113) = 1.02$, $p = .3149$, $f = .095$. Earlier work (e.g., Hugenberg, 2005) could not detect the greater influence of threat than negative-valence for the perception of Black than White faces because angry and sad faces were not previously paired together and each only co-occurred with happy faces.

Unlike Studies 1 and 2, this study did not find that White faces were associated more strongly than Black faces with positive valence. Notice, in particular, that with the Calm distractor, time to begin categorizing happy faces as Cheerful did not differ among the races (middle row of Figure 7). Likewise, Cheerful did not differentially distract across the races from time to begin categorizing neutral faces as Calm (middle row of Figure 8). Some might wonder whether a White-positive (and Asian-positive) association manifested in the earlier times to begin categorizing happy faces as Cheerful for White (and Asian) than Black faces when the distractor was Dangerous (top row of Figure 7) or Depressed (bottom row of Figure 7). If such were the case, however, time to begin categorizing happy faces as Cheerful with a Calm distractor should have occurred earlier when they were White (and Asian) than Black, just as time to begin categorizing angry faces as Dangerous with a Calm distractor occurred earlier in time when they were Black than White (and Asian; middle row Figure 5). Indeed, Black faces were processed earlier in time (than White and Asian faces) as a threat. But that earlier processing bias did not manifest with positivity for White (or Asian) faces. So why did the current study not find evidence for a stronger White than Black association with positivity? One possibility is our neutral label. Perhaps Calm (like Cheerful) is a positive, rather than neutral,

attribute. If so, Cheerful and Calm would each be appropriate for a smiling White face (if White individuals are associated with positivity) and inappropriate for a smiling Black face (if Black individuals are not associated with positivity). When paired together on trials, Cheerful and Calm would no longer be uniquely diagnostic and perhaps that is why we were unable to observer a stronger White than Black association with positivity. Indeed, subsequent pilot-testing (N = 22) indicated that "calm" was rated more positively (M = 4.81, SD = 1.63 on a 1 – 7 scale) than each of 12 other ostensibly neutral words (e.g., common, generic, neutral, plain; Ms < 2.05, SDs < 1.47). Study 4 addressed this issue by using different labels.

That Calm is a positive (rather than neutral) attribute also provides an alternative account for the evidence of the Black-negative association in the current study. Evidence for a Black-negative association is provided by the findings that White participants were (a) faster to begin categorizing sad Black than sad White or Asian faces as Depressed when Cheerful was a distractor and (b) slower to begin categorizing neutral Black than neutral White or Asian faces as Calm when Depressed was a distractor. Because categorization can be facilitated by the target label and inhibited by the distractor label, it is possible that the latter patterns reflect a White-(and Asian-) positive association rather than a Black-negative association. In particular, (a) categorization of sad White and Asian faces could have been inhibited by positivity via the Cheerful distractor-label rather than categorization of sad Black faces being facilitated by negativity via the Depressed target-label and (b) categorization of neutral White and Asian faces could have been facilitated by positivity via the Calm target-label rather than categorization of sad Black faces being inhibited by negativity via the Depressed distractor-label. This possibility is strengthened by the aforementioned post-test data indicating Calm may have been a more positive target than intended. To better distinguish the facilitating effect of the target-label from

the inhibiting effect of distractor label in Study 4, we changed the label for neutral faces from Calm to the negation of the label for the emotive faces with which they are paired within blocks (e.g., Not-Dangerous when neutral faces occur with angry faces).

Study 4

Study 4 uses mouse-tracking to test the association of Black vs. White males with threat, negativity, and positivity, and directly tests whether Black-vs-White-threat is stronger than Black-vs-White-negative via trials that pit threat against negativity. We did not use Asian faces because participants in the previous study responded similarly to Asian and White faces. We retained the same Black and White angry, sad, happy, and neutral faces, and retained "Dangerous" as the target label for angry faces. We changed the target label for sad and happy faces, respectively to "Negative," and "Positive." We used as the target label for neutral faces, which were paired within blocks with one of the emotive faces (i.e., angry, sad, or happy), the negation of the label for the emotive face: (a) "Not-Dangerous" when paired with angry faces, (b) "Not-Negative" when paired with sad faces, and (c) "Not-Positive" when paired with happy faces. Three blocks separately presented angry and neutral faces with the labels "Dangerous" and "Not-Dangerous," sad and neutral faces with the labels "Negative" and "Not-Negative," and happy and neutral faces with the labels "Positive" and "Not-Positive." Hence, for emotive faces we used the negation of the target-label as the distractor-label (e.g., Not-Dangerous, Not-Negative, Not-Positive) to better isolate the influence of the target from the inhibiting effect of the distractor. Nonetheless, to again assess the relative association of threat and negativity when they are simultaneously paired, we presented in a fourth block angry and sad faces with the labels "Dangerous" and "Negative."

If White Americans associate Whites more than Blacks with positivity, participants in the

happy-neutral block should begin moving earlier in time to Positive for happy White than Black faces and later in time to Not-Positive for neutral White than Black faces. If White Americans associate Blacks more than Whites with negativity, participants in the sad-neutral block should begin moving earlier in time to Negative for sad Black than White faces and later in time to Not-Negative for neutral Black than White faces. If White Americans associate Blacks more than Whites with threat, participants in the angry-neutral block should begin moving earlier in time to Dangerous for angry Black than White faces and later in time to Not-Dangerous for neutral Black than White faces. If, as Study 3 indicates, White Americans associate Blacks more strongly than Whites with threat than negativity, participants in the angry-sad block should begin moving earlier in time to Dangerous for angry Black than White faces but not earlier in time to Negative for sad Black than White faces. That is, threat will distract from negativity to delay the categorization of a sad Black face as negative, but negativity will not distract from threat. If, however, threat does not distract from negativity, participants will begin moving earlier in time to both Dangerous and Negative, respectively, for angry and sad Black than White faces. We determined sample size by the number of participants we could run in a semester.

Methods

White undergraduates (N = 327, 217 females) participated for partial credit in an introductory psychology course and sat in separate cubicles with a 48cm high-speed, high-resolution monitor and computer. Instructions explained that future studies require face pictures that can be quickly and accurately identified as positive, negative, or dangerous. Those labels were defined such that positive faces “look happy, friendly, joyful,” negative faces “look sad, gloomy, unhappy,” and dangerous faces “look angry, scary, threatening.” They were also told that they would identify emotionless-neutral faces that are “not positive, not negative, and not

dangerous." Participants were shown examples of Black and White faces displaying each expression identified by each label. The remaining instructions were the same as Study 3, and participants did the same set of fruit-vs-vegetable practice trials before completing four blocks (40 trials each) of 160 face-trials and being debriefed.

Each of the first three blocks presented neutral faces with one of three emotive faces (i.e., angry, sad, happy). The order of those blocks – i.e., angry and neutral faces (with Dangerous and Not-Dangerous as labels), sad and neutral faces (with Negative and Not-Negative as labels), and happy and neutral faces (with Positive and Not-Positive as labels) – was randomized. The fourth block presented angry and sad faces with Dangerous and Negative as labels. On any trial, one label (target) described the expression, and the other label (distractor) did not. Within a block, 20 Black and 20 White faces displayed each expression, and the order of faces and left-right starting position of labels was randomized (with the left-right label positions switching halfway through the block).

Data were recorded in MouseTracker and exported to SAS as in Study 3. Of the 52,320 trials, 16 (0.612%) were not recorded by MouseTracker, 732 (1.40%) did not have recorded data because the participant exceeded the 2000ms limit, and 1,066 (2.04%) ended incorrectly (i.e., participant clicked the distractor, not target, label). Of the 50,506 useable trials, based on our a priori criteria, we excluded 2,062 (4.08%) in which movement initiated too late (>300ms of clicking start) and 922 (1.83%) that ended abnormally fast (<600ms), yielding 47,522 trials from 327 participants.

As in Study 3, we calculated each participant's TICC for each facial expression of each race (March & Gaertner, 2021). For stable mouse-trajectory estimates, we averaged the time-synched x-, y-coordinates across the (up to) 10 trials of each Facial-expression x Race pairing

within a block (e.g., 10 trials of angry Black-faces to Dangerous when paired with neutral-faces) for each participant. This yielded for each participant 16 average trajectories (4 from each block). For each of those average trajectories, we (a) calculated the Euclidean distance of the mouse at each time interval from the target and distractor labels, respectively, and (b) estimated the corresponding TICC (i.e., point in time when participant began moving relatively closer to the target than distractor label; the non-linear Gompertz and Baranyi models for TICC both converged on 99.23% of the 5,228 curves with an average pseudo-R^2 = .9373).

Results

To test whether White Americans are biased earlier in the decision process for Black than White faces by positivity, threat, and/or negativity, we entered participants' TICC estimate for a given block in into a 2(Race: Black vs. White) x 2(Within-block Face: Happy vs. Neutral; Angry vs. Neutral; Sad vs. Neutral; Angry vs. Sad) multivariate repeated-measures ANOVA (degrees of freedom vary due to missing TICC estimates from non-convergence or missing trajectories). Figure 9 displays the average TICC to the target label for each Race x Face pairing for each block.

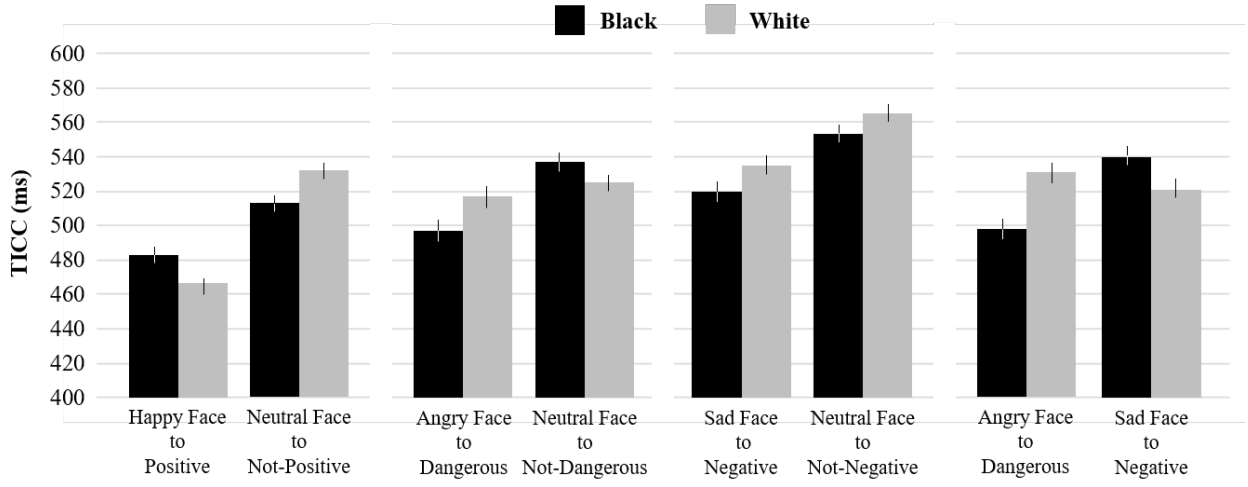


Figure 9. Mean TICC to target label for each Race x Face pairing in the four blocks. Error bars are ±1 *SEM* calculated within-participants (O'Brien & Cousineau, 2014).

Positive / Not-Positive for happy and neutral faces. TICC to the target label was influenced by a Race x Face interaction, $F(1, 319) = 16.75$, $p < .0001$, $f = .23$, indicating a stronger positivity association with White than Black faces. Participants began moving *earlier* in time to Positive when the happy face was White (M = 466ms) than Black (M = 483ms), $F(1, 319) = 8.22$, $p = .0044$, $d_z = 0.16$, and *later* in time to Not-Positive when the neutral face was White (M = 532ms) than Black (M = 513ms), $F(1, 319) = 9.80$, $p = .0019$, $d_z = 0.17$. Decomposed within levels of race, the interaction also indicates that the tendency to begin moving earlier in time to Positive for happy faces than to Not-Positive for neutral faces was stronger when faces were White, $F(1, 319) = 115.90$, $p < .0001$, $d_z = 0.60$, than Black, $F(1, 319) = 21.44$, $p < .0001$, $d_z = 0.23$. Consistent with Studies 1 and 2, these data indicate that White participants more strongly associate positivity with White than Black males.

Dangerous / Not-Dangerous for angry and neutral faces. TICC to the target label was influenced by a Race x Face interaction, $F(1, 318) = 7.77$, $p = .0056$, $f = .17$, indicating a stronger threat association with Black than White faces. Participants began moving *earlier* in time to Dangerous when the angry face was Black (M = 497ms) than White (M = 517ms), $F(1, 318) = 4.74$, $p = .0301$, $d_z = 0.12$, and *later* in time to Not-Dangerous when the neutral face was Black (M = 537ms) than White (M = 525ms), $F(1, 318) = 3.23$, $p = .0734$, $d_z = 0.08$. Decomposed within levels of race, the interaction also indicates that the tendency to begin moving earlier in time to Dangerous for angry faces than to Not-Dangerous for neutral faces occurred when faces were Black, $F(1, 318) = 21.12$, $p < .0001$, $d_z = 0.24$, but not White, $F(1, 318) = 0.91$, $p = .3420$, $d_z = 0.06$. Consistent with Studies 1 2, and 3, these data indicate that White participants more strongly associate threat with Black than White males.

Negative / Not-Negative for sad and neutral faces. TICC to the target label was not

influenced by a Race x Face interaction, $F(1, 308) = 0.07$, $p = .7906$, $f = .02$, indicating that negativity is not differentially associated with Blacks and Whites. Instead, lag-time was affected by two main effects. Regardless of the race of the face, participants began moving earlier in time to Negative for sad faces (M = 527ms) than to Not-Negative for neutral faces (M = 559ms), $F(1, 308) = 20.95$, $p < .0001$, $f = .26$. Regardless of the emotion of the face, participants began moving to the target label earlier in time for Black (M = 537ms) than White faces (M = 550ms), $F(1, 308) = 5.43$, $p = .0209$, $f = 0.13$. Consistent with Studies 1 and 2, these data indicate that White participants do not differentially associate nonthreatening negativity with Black or White males.

Dangerous / Negative for angry and sad faces. TICC to the target label was influenced by a Race x Face interaction, $F(1, 315) = 19.03$, $p < .0001$, $f = .25$, indicating that Blacks (relative to Whites) are more strongly associated with threat than negativity. Participants began moving *earlier* in time to Dangerous when the angry face was Black (M = 498ms) than White (M = 531ms), $F(1, 315) = 19.36$, $p < .0001$, $d_z = 0.25$ and *later* in time to Negative when the sad face was Black (M = 539ms) than White (M = 521ms), $F(1, 316) = 4.35$, $p = .0378$, $d_z = 0.13$. Decomposed within levels of race, the interaction also indicates that the tendency to begin moving earlier in time to Dangerous for angry faces than to Negative for sad faces occurred when faces were Black, $F(1, 315) = 25.19$, $p < .0001$, $d_z = 0.28$, but not White, $F(1, 315) = 1.45$, $p = .2287$, $d_z = 0.09$. That participants were (a) faster to begin categorizing as Dangerous angry Black than White faces, but (b) not faster to begin categorizing as Negative sad Black than White faces in a situation in which Dangerous and Negative co-occurred as response labels suggests that the threat association is stronger than the negative association for the perception of Black (than White) faces.

Indeed, the relative influence of threat over negativity in the processing of Black faces

can be appreciated by comparing the race effect on TICCs for angry and sad faces in the last three panels of Figure 9. In both the angry-neutral and angry-sad blocks, participants began earlier in time to categorize as Dangerous angry Black than angry White faces. That is, TICC to Dangerous occurred earlier for angry Black than White faces regardless of whether the distractor label was Not-Dangerous or Negative. TICC to Negative, in contrast, did not evidence such consistency. In particular, TICC to Negative was (a) *earlier* for sad Black than White faces when the distractor was Not-Negative (i.e., sad-neutral block), but (b) *later* for sad Black than White faces when the distractor was Dangerous (i.e., sad-angry block). In the latter case, the option of categorizing a sad Black face as Dangerous delayed the time at which participants began moving toward the Negative label. Thus, as in Study 3, there is evidence of a Black-threat association that distracted participants from categorizing Black faces as negative, but not evidence of a Black-negative association that distracted participants from categorizing Black faces as dangerous.

Discussion

Consistent with Studies 1, 2, and 3, this study indicates that threat is associated more strongly with Blacks than Whites. Consistent with Studies 1 and 2, but not 3, this study indicates that positivity is more strongly associated with Whites than Blacks. This suggests that the lack of evidence for a White-positive association in Study 3 was likely due to our use of Calm as a target label for neutral faces (and a distractor for emotive faces). As noted when discussing Study 3, calm is a positive attribute, as is Cheerful (i.e., the label for happy faces), and the simultaneous pairing of Calm and Cheerful for neutral and happy faces likely distracted from the categorization of happy White faces as Cheerful and undermined our ability to detect a stronger positive association with Whites than Blacks. With altered target labels in the current study, we

observed a White-positive association.

This also suggests that the Study 3 evidence of a Black-negative association was likely an artifact of the White-positive association via the facilitating effect of Calm as a target-label for White neutral faces rather than the distracting effect of Depressed on neutral Black faces. Indeed, consistent with Studies 1 and 2, the sad-neutral block in the current study found no evidence of a Black-negative association such that race had no differential effect on the time at which participants began categorizing sad and neutral faces as Negative vs. Not-Negative.

Finally, consistent with Study 3, this study found evidence of a Black-threat association that distracted from categorizing Blacks in terms of negativity but not a corresponding Black-negative association that distracted from categorizing Blacks as threatening. This indicates that the Black-threat association is stronger than a Black-negative association.

Study 5

Study 5 tests whether the Black-vs-White threat association is stronger than the Black-vs-White negative association by combining the evaluative priming task of Studies 1 and 2 with the simultaneous pairing of threat and negativity of Studies 3 and 4. Furthermore, we use new stimuli to address alternative explanations suggested by anonymous reviewers.

Studies 1 and 2 used images as primes and targets. Perhaps the results were due to variation in the lightness of the images, rather than a Black-threat association, such that the darkness of the Black (relative to White) faces facilitated the possibly darker threatening images than lighter positive images. Inconsistent with such a possibility is that the target sets do not differ in luminance (M_{threat} = 123.03, M_{posivite} = 126.32, M_{negative} = 125.17). Nonetheless, we rule out the role of differential luminance by using words rather than images as stimuli.

Studies 3 and 4 required participants to categorize emotive faces. Perhaps participants

began categorizing angry Black faces as dangerous earlier than angry White or Asian faces because of a stronger stereotype of anger for Black (than White or Asian) men (i.e., they expect Black men to be angry). There are, however, three counter-points. First, a stronger anger stereotype for Black men could be a consequence of a Black-threat association (though the stereotype could reflect an understanding of how Black men are treated; i.e., they have reason to be angry). Second, an anger-stereotype account cannot parsimoniously explain why neutral Black-faces facilitated evaluation of threatening images (e.g., scorpions, fire, wolves, gunmen) more than neutral White-faces in Studies 1 and 2. Third, in Studies 3 and 4 the dangerous label not only facilitated categorization of angry Black faces, but it also distracted categorization of neutral, sad, and happy Black faces. Said otherwise, the patterns were not simply a product of angry Black-faces – all Black faces were affected by the Dangerous label. Nonetheless, the stimuli in Study 5 are devoid of angry faces and emotion referents.

Two pilot studies (see supplemental materials) identified words that manipulate race and distinguish threat from negativity. They yielded 12 first-names considered typical of either Black men (Darnell, DeAndre, DeShawn, Jamal, Tyrone, Trevon) or White men (Brad, Connor Ethan, Jack, Jake, Scott) and 12 negatively-valenced words that denote either physical threat (aggressive, harmful, murderous, threatening, unsafe, violent) or nonthreatening negativity (awful, disliked, displeasing, inferior, lousy, undesirable). Participants in Study 5 experienced a 2(Prime name: Black vs. White) x 2(Target word: Threat vs. Nonthreatening negative) within-subjects design in which names preceded target words and they indicated whether the target was “dangerous” or “negative.” The absence of positive targets (unlike Studies 1 and 2) allowed separate responses for threatening and negative targets (rather than yoking them to a common response, “bad”). We determined sample size by the number of participants obtained by the end

of the semester.

Methods

White undergraduates (N = 206; 156 females) at a southeastern university participated online for partial credit in an introductory psychology course. We administered the study with Inquisit Web (https://www.millisecond.com), which maintains millisecond accuracy by installing an app on the user's computer (Windows or Mac). Instructions explained that pairs of words would be presented sequentially with the first being a name and the second a target, and the task is to indicate as quickly and accurately as possible whether the target is negative or dangerous (by pressing the "A" or "L" key, respectively). Participants practiced 24 trials involving only target words and transitioned with a button click to complete 144 trials before being debriefed. Each of the 144 trials began with a centrally located string of asterisks ("******") for 400ms that functioned as a fixation, which was replaced for 400ms by a name, which was replaced for 400ms by a target word, and ended on response to the prompt of whether the target was Negative or Dangerous. A 1500ms blank screen separated trials. The order and pairing of prime and target were randomized with all targets presented once before any was represented.

Response latency to a correct response (i.e., negative for negative targets, dangerous for threatening targets) was the dependent measure. Two participants ended the study after one trial. One threatening target ("unsafe") had error rates above 20% (our a priori cutoff). We excluded responses to that error-prone target (n = 2,448, 8.3%) and all remaining incorrect responses (n = 1,756, 6.5%) leaving 25,172 correct responses. We subsequently excluded slow times exceeding three interquartile ranges of the 75th percentile (Tukey, 1977; n = 701, 2.78%), and nine participants with less than 70% of their data remaining, yielding 195 participants (23,748 responses).

Results

Before testing the hypothesis, we examined whether reaction time was affected by either the length (i.e., number of letters) of the prime names or target words. Black names were longer than White names (*M*s = 6.33 vs 4.67 letters). Threatening words were approximately the same length as negative words (*Ms* = 8.33 vs 8.00 letters). In Proc Mixed of SAS, we regressed natural logged reaction times on name-length and target-length with random effects of the intercept, name-length, target-length, and their covariances for participants, and a random intercept for targets (the model would not converge with a random intercept for primes). Reaction times were unrelated to the length of target words, $F(1, 194) = 1.13$, $p = .2901$, but significantly delayed by the length of prime names, $F(1, 194) = 20.40$, $p = .0001$, such that each additional letter of a name increased reaction time by approximately 2.9 ms. Consequently, to unconfound the race of the prime-name from the length of the prime-name, we control name-length in subsequent analyses.

With Proc Mixed of SAS, we regressed natural-logged reaction times on a 2(Prime: Black, White) x 2(Target: Threatening, Negative) factorial with name-length as a covariate, random effects of the intercept, target, Prime x Target, and their covariances for participants, and a random intercept for target-stimuli. (Models would not converge with a random prime effect for participants or a random intercept for prime-stimuli.)[8] Independent of the delaying effect of name-length, $F(1, 23152) = 13.27$, $p = .0003$, and consistent with a stronger Black than White association with threat than negativity was the Black vs. White x Threatening vs. Negative interaction, $F(1, 194) = 17.79$, $p < .0001$, $d_0 = -0.13$. As displayed in Figure 10, Black (versus White) names yielded a faster response to threatening targets, $F(1, 23152) = 11.80$, $p = .0006$, $d_0 = -0.08$, ($M_{black} = 621$ ms vs. $M_{white} = 632$ ms) and a non-significantly slower response to

negative targets, $F(1, 23152) = 3.28$, $p = .0700$, $d_0 = 0.04$, (M_{black} = 668 ms vs. M_{white} = 664 ms). Decomposed within levels of race, the interaction also indicates that threatening (versus negative) targets yielded faster responses when primed by Black names, $F(1, 194) = 4.94$, $p = .0275$, $d_0 = -0.37$, but not White names, $F(1, 194) = 2.14$, $p = .1449$, $d_0 = -0.24$. These data indicate that White Americans more strongly associate Black than White men with threat than negativity.[9]

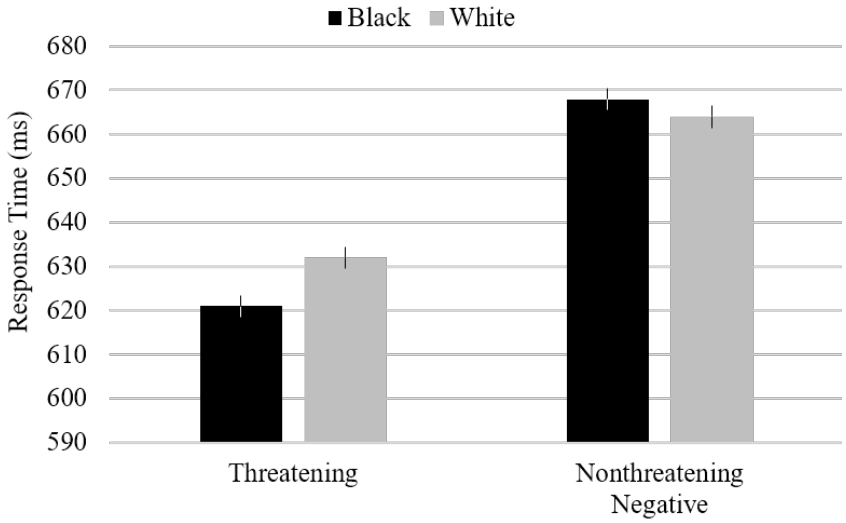


Figure 10. Mean response time as a function of Prime and Target. Error bars are ±1 *SEM* calculated from clustered multilevel data (Gelman & Hill, 2006).

Discussion

This study combined the evaluative priming task of Studies 1 and 2 and the simultaneous pairings of threat and negativity of Studies 3 and 4. The stimuli, however, consisted of names and words rather than images and emotional faces and rules-out alternative explanations regarding luminance or emotion stereotypes. The results conceptually replicate the race effects on threat versus negativity of Studies 1 and 2 and Studies 3 and 4. In particular, Black (relative to White) names facilitated responding to threatening words as “Dangerous” but not nonthreatening negative words as “Negative.” These data indicate that White Americans more strongly associate

Black than White men with physical threat than negativity.

General Discussion

The Dual Implicit Process Model distinguishes the implicit processing of threat (i.e., can it harm/kill me?) and valence (i.e., do I like/dislike it?). From the vantage of this model, responses to social groups can be driven by threat, valence, or both. Whether a group is associated with physical threat can have drastically different consequences than if it is associated only with valence (negative or positive). When we examined the research literature on implicit anti-Black bias, however, we noticed two limitations that prevent conceptual clarity. With few exceptions, there was no methodological distinction between threat and negativity despite the fact that threatening stimuli are also negative. Consequently, it is not clear whether automatic anti-Black bias in past research reflects threat, negative valence, or both. Also, with few exceptions, there was no consideration of other racial outgroups to assess whether White's implicit anti-Black bias via threat or negativity is unique to Blacks or a more general intergroup phenomenon. The purpose of the current work was to address those limitations to assess whether White Americans' associate Black men with threat, negativity, or both.

We conducted five studies that independently operationalized threat and negativity to enable their differentiation. All five studies unambiguously indicated that White Americans automatically evaluate Black men as a survival threat. Study 3 indicated that association was unique to Black men and did not extend to Asian men. Finally, Studies 3, 4, and 5, which simultaneously paired threat against negativity, indicated that White Americans more strongly associate Black (relative to White or Asian) men with threat than negativity.

The take-home message of our findings should be contextualized within the methods that utilized speeded tasks and examined reactions early in the decision process. The findings

should not be interpreted as suggesting that White Americans lack negative (or positive) stereotypes of Black men or that all implicit bias is threat based. The primary finding is that White American's initial (i.e., early, or automatic) evaluation of Black men is that they pose a survival threat. To better contextualize this research, we subsequently integrate our findings with the broader bias literature from the perspective of the DIPM.

Situating in the Broader Literature

We do not claim that threat is the only association White Americans have regarding Black men, but we do claim that it is often the initial evaluation. The threat association is critical in regard to the DIPM because it implies that Black men could activate in White Americans implicit threat-processing that initiates rapid threat-responses geared toward self-preservation. We think this model provides an important insight, in particular, into why police use greater force in encounters with Black Americans than with other races (Goff et al., 2016). Rather than reflecting implicit dislike or disdain, we suspect such shootings reflect threat responses via implicit threat evaluation. Of course, not every instance of undue force is a product of implicit threat processing; more deliberate and delayed decisions for force could certainly be a product of disdain (such as applying a choke-hold or knee-to-the-neck of an already subdued suspect).

From a DIPM perspective, additional automatic valence and explicit responses can occur as well, consistent with a wealth of the prejudice literature. Like threat responses, these automatic valence and explicit responses stem from previous learning about Black Americans, and there is no shortage of it (Ruscher, 2001). From an early age, children are confronted with negative depictions and messages regarding Black Americans from media, peers, parents, and authority figures. These clearly have impact, as investigations of the racial prejudices of young

children attest. For example, White children as young as 6 years old exhibit automatic bias against Black Americans, and by 10 that bias is on par with that of adults (e.g., Baron & Banaji, 2006). Although some of this learning entails the acquisition of Black-threat associations, the content of these automatic biases can be independent of physical threat (e.g., "unintelligent," "lazy"), and can be found both in children (Cvencek et al., 2015) and adults (Devine, 1989; Wittenbrink et al., 2001). From the perspective of the DIPM, as they are not threat-relevant but can still exhibit automatic properties, such associations are considered distinct and downstream from threat responses.

Explicit, deliberatively-held beliefs about Black Americans can recapitulate negative automatic stereotypes. Examples can be found within explicit measures of prejudice like the Symbolic Racism 2000 scale (Henry & Sears, 2002), which include negative beliefs about Black Americans (e.g., "if blacks would only try harder, they could be just as well off as whites") that many White Americans endorse (Olson & Zabel, 2016). Explicitly-held beliefs about Black Americans can also be positive, and often entail an appreciation for Black cultural contributions and their overcoming of social and institutional obstacles (Czopp & Monteith, 2006). Such positive explicitly-held beliefs are often at odds with automatic responses and can support motives to redress inequities and treat Black individuals positively (Dunton & Fazio, 1997; Plant & Devine, 1998).

In short, a threat association is not the only association White Americans have of Black Americans. However, from the DIPM perspective, as an initial process, threat has important implications for later automatic valence processing (entailing other valenced responses, including nonthreat-oriented stereotypes with automatic properties) and explicit processing (entailing deliberately-held beliefs and motives). Specifically, the DIPM proposes that threat

processing can potentiate valence and later explicit processes. For example, a White American's automatic threat evaluation of a Black individual may increase the likelihood of perceiving other negative stereotypic traits in that individual. Such automatic processes—both threat and later automatic valenced processes—can affect perception, judgment, and behavior toward Black Americans, particularly when there is limited opportunity and motivation to act otherwise (Fazio & Olson, 2014).

The necessity for time, cognitive capacity, and motivation to counteract automatic threat and valence responses has clear implications for when these responses are likely to predominate. For example, in situations where a rapid response is required, threat is likely to be the dominant one. From the DIPM perspective, such situations are not limited to survival-oriented contexts (e.g., the shooter-bias paradigm), and apply to any rapidly rendered perception or judgment (e.g., impressions formed very quickly). Given more time, other valence-related information, including stereotypes with automatic properties that are not threat-related, may affect perceptions. Thus, The DIPM proposes that White American's very rapid response to a Black American will be more threat-based, and later but still relatively automatically-formed responses will start to incorporate other valenced knowledge. Finally, explicit responses will mirror automatic ones when there is little motivation or opportunity to do otherwise. If motivated and able, individuals can deliberate on motives, values, and contextual information to correct for their automatic responses. For example, in the domain of political decision making, beliefs about the historical plight of Black Americans or notions about Black individuals violating the protestant work ethic will enter into judgments, decisions, and behavior (e.g., Biernat et al., 1996). But ultimately, according to the DIPM, the series of processes that end in some perception, judgment, emotion, or behavior, begins with an automatic evaluative focus on

whether the target poses a survival threat. The findings we report here suggest that to White Americans, Black Americans do just this.

Reducing Anti-Black Bias via the DIPM

Empirical work unfortunately indicates that implicit-bias interventions have little to no lasting impact (Lai et al., 2016). Interventions take many forms. Some use counter-stereotypic exemplars such as having participants imagine being attacked by a White assailant and rescued by a Black hero or practice a partial IAT in which Black was paired with Good (and positive Black exemplars, e.g., Oprah) and White was paired with Bad (and negative White exemplars, e.g., Hitler). Others use evaluative (re)-conditioning, which involves repeated pairings of Black faces with positive words and White faces with negative words. Given the findings of the current research and the threat-valence distinction of the DIPM, it seems that interventions should directly target the Black-threat association. Persistent bias reduction (i.e., actual change) may not result simply from increasing a White-threat association or decreasing a Black-negative association and increasing a Black-positive association. Instead, more effective and lasting interventions may result from reducing the automatic association between Black and "danger."

Most prejudice reduction research is situated within the contact-hypothesis (Pettigrew & Tropp, 2013), and perhaps repeated cooperative interactions could help to undo the Black-threat association. However, research on fear conditioning suggests that reducing the Black-threat association is likely to be difficult (Hermans et al., 2006). Contact may operate analogously to phobic exposure therapy for reducing threat associations and threat responses, but, informed by the DIPM and exposure therapy research, it might be most effective when introduced incrementally. For example, during phobia interventions, individuals experience incrementally increased exposure to a threat stimulus (e.g., imagine a snake □ look at a picture of a snake □

look at a live snake hold a live snake; Hofmann, 2008). The result of repeated safe exposure to the stimulus is a reduced threat response. Nonetheless, without targeting the Black-threat association, anti-Black bias interventions are likely to remain ineffective.

Threat Associations and Threat Responses

Readers might question whether the descriptively faster responses to positive targets in Figure 1 and happy faces in Figure 9 than to their threatening counterparts (threatening targets, angry faces) is inconsistent with the proposition of the DIPM that implicit threat processing precedes implicit valence processing. We remind the reader that the faster and stronger responses proposed by the DIPM are threat responses geared toward self-preservation via rapid detection and avoidance of immediate danger (e.g., LeDoux, 2012; Öhman & Mineka, 2001; Vuilleumier, 2005). Such effects manifest as stronger/earlier responses of the amygdala (Kveraga et al., 2015; Méndez-Bertolo et al., 2016), skin-conductance (Knight et al., 2009), startle-eyeblinks (March et al., 2017), ocular movements to the threatening stimulus (Hermans et al., 1999; March et al, 2017), earlier detection of the threat (Blanchette, 2006; March et al, 2017), and reflexive freezing and fighting (LeDoux, 2014; Löw et al. 2015). Button presses and mouse movements in our evaluative priming and mouse-tracking tasks are not self-preserving threat-responses, they are measures of associations.

Faster latencies to positive targets, particularly in evaluative priming, have been documented (Unkelback et al., 2008). To the extent to which such button presses are irrelevant to the threat response, they may be delayed to threatening targets as relevant responses unfold. We utilized the button presses and mouse movements to assess threat and valence associations as a function of race, not to assess threat responses to race. There is, however, evidence consistent with the possibility that White Americans evidence a threat response to Black males

(Amodio, 2014). For example, (a) they evidence stronger amygdala activation to Black than White male faces at presentation rates of 30ms (Cunningham et al., 2004) and 2000ms (Phelps et al. 2000; but, inconsistently, no difference at 500ms (Cunningham et al., 2004)), (b) their amygdala activation is particularly strong when Black faces have a direct (but not averted) eye-gaze (Richeson, Todd, Trawalter, & Baird, 2008), and (c) they have stronger startle-eyeblink responses to Black than White or Asian male-faces (Amodio et al., 2003). Finally, as we explained previously, police officers' differential use of force toward Black suspects could be understood as a threat response. Of course, the threat response requires that the perceiver associate Black men with threat and the current work indicates that such is the case for White Americans.

Limitations of the Current Work

Perhaps our relative lack of evidence of a Black-negative association has something to do with our operationalization of negativity. In Studies 1 and 2 we used previously validated image sets that distinguish threat and negativity. In line with earlier work (Donders et al., 2008; Judd et al., 2004; Todd et al., 2016), the negative set consisted of images that evoke disgust (such as insects, excrement, and decayed teeth) or sadness (such as injured kitten, dead dogs). The Black vs. White prime effect did not differ as a function of whether the negative targets were disgust related or sadness related (i.e., Black vs. White x Disgust vs. Sad), $F_{study1}(1, 78) = 1.84$, $p = .1785$, $f = .154$ and $F_{study2}(1, 131) = 2{,}42$, $p = .1218$, $f = .136$. In Studies 3 and 4 we used sad faces and response labels of Depressed (Study 3) or Negative (Study 4). Unlike evidence of a Black-threat association that occurred in every study, evidence of a Black-negative association occurred only in Study 3 and it was rivaled by the threat association on paired trials. In Study 5, we used yet another operationalization of negativity (i.e., negatively

valenced words that are unrelated to threat) and again found no evidence of a Black-negative association but we continued to find a Black-threat association. Had we operationalized negativity in a way that emphasized negative non-threatening Black male stereotypes, perhaps we would have found evidence of a Black-negative association (March et al., 2020).

We focused in the current work on perceptions of Black men. Whether White Americans additionally associate physical threat with Black women is an empirical question. There is evidence that conditioned outgroup fear is slower to extinguish when it is conditioned on male than female exemplars (Navarrete et al., 2009). However, the tendency for Black (vs. White) faces to facilitate the identification of guns is not limited to the faces of Black men and similarly occurs when the faces are of Black boys, Black women, and Black girls (Thiem et al., 2019).

Conclusion

Informed by the DIPM (March et al., 2018a, b), we methodologically differentiated threat from valence to understand White American's initial perceptions of Black men. The data indicate that White American's initial (i.e., early or automatic) evaluation of Black men is that they pose a survival threat.

References

Al-Janabi, S., MacLeod, C., & Rhodes, G. (2012). Non-threatening other-race faces capture visual attention: Evidence from a dot-probe task. *PLOS ONE, 7*, 1-7.

Amodio, D. M., Harmon-Jones, E., & Devine, P. G. (2003). Individual differences in the activation and control of affective race bias as assessed by startle eyeblink response and self-report. *Journal of personality and social psychology*, *84*, 738-753.

Amodio, D. M., & Ratner, K. G. (2011). A memory systems model of implicit social cognition. *Current Directions in Psychological Science*, *20*, 143-148.

Amodio, D. M. (2014). The neuroscience of prejudice and stereotyping. *Nature Reviews Neuroscience*, *15*, 670-682.

Amodio, D. M. (2019). Social Cognition 2.0: An interactive memory systems account. *Trends in Cognitive Sciences*, *23*, 21-33.

Baron, A. S., & Banaji, M. R. (2006). The development of implicit attitudes: Evidence of race evaluations from ages 6 and 10 and adulthood. *Psychological science*, *17*(1), 53-58.

Baty, F., & Delignette-Muller, M.-L. (2004). Estimating the bacterial lag time: which model, which precision? *International Journal of Food Microbiology, 91*, 261-277.

Biernat, M., Vescio, T. K., & Theno, S. A. (1996). Violating American values: A "value congruence" approach to understanding outgroup attitudes. *Journal of Experimental Social Psychology*, *32*(4), 387-410.

Blanchette, I. (2006). Snakes, spiders, guns, and syringes: How specific are evolutionary constraints on the detection of threatening stimuli? *The Quarterly Journal of Experimental Psychology*, *59*, 1484-1504.

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, 6*, 1314-1329.

Correll, J., Urland, G. L., & Ito, T. A. (2006). Event-related potentials and the decision to shoot: The role of threat perception and cognitive control. *Journal of Experimental Social Psychology, 42,* 120-128.

Cottrell, C.A., & Neuberg, S.L. (2005). Different emotional reactions to different social groups: A sociofunctional threat-based approach to "prejudice." *Journal of Personality and Social Psychology, 88*, 770-789.

Cunningham, W.A., Johnson, M. K., Raye, C.L., Gatenby, C., Gore, J.C., & Banaji, M.R. (2004). Separable neural components in the processing of Black and White faces. *Psychological Science, 15,* 806-813.

Cvencek, D., Nasir, N. I. S., O'Connor, K., Wischnia, S., & Meltzoff, A. N. (2015). The development of math–race stereotypes: "They say Chinese people are the best at math". *Journal of Research on Adolescence*, *25*(4), 630-637.

Czopp, A. M., & Monteith, M. J. (2006). Thinking well of African Americans: Measuring complimentary stereotypes and negative prejudice. *Basic and Applied Social Psychology*, *28*(3), 233-250.

Devine, P. G. (1989). Stereotypes and prejudice: Their automatic and controlled components. *Journal of personality and social psychology*, *56*(1), 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.

Donders, N. C., Correll, J., & Wittenbrink, B. (2008). Danger stereotypes predict racially biased attentional allocation. *Journal of Experimental Social Psychology, 44,* 1328-1333.

Dovidio, J., Kawakami, K., Johnson, C., Johnson, B., & Howard, A. (1997). The nature of prejudice: Automatic and controlled processes. *Journal of Experimental Social Psychology, 33,* 510–540.

Duncan, B. L. (1976). Differential social perception and attribution of intergroup violence: Testing the lower limits of stereotyping of Blacks. *Journal of Personality and Social Psychology, 34,* 590–598.

Dunton, B. C., & Fazio, R. H. (1997). An individual difference measure of motivation to control prejudiced reactions. *Personality and Social Psychology Bulletin, 23*(3), 316-326.

Eberhardt, J.L., Goff, P.A., Purdie, V.J., & Davies, P.G. (2004). Seeing black: Race, crime, and visual processing. *Journal of Personality and Social Psychology, 87*, 876-893.

Ekman, P., & Friesen, W. (1978). *Facial Action Coding System: A technique for the measurement of facial movement*. Palo Alto: Consulting Psychologists Press.

Fazio, R. H. (1990). Multiple processes by which attitudes guide behavior: The MODE model as an integrative framework. *Advances in Experimental Social Psychology, 23,* 75–109.

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.

Fazio, R. H., & Olson, M. A. (2014). The MODE model: Attitude-Behavior Processes as a Function of Motivation and Opportunity. In Sherman, J. W., Gawronski, B., & Trope, Y. (Eds.). *Dual process theories of the social mind.* New York: Guilford Press.

Freeman, J. B. (2018). Doing psychological science by hand. *Current Directions in Psychological Science, 27,* 315-323.

Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

Goff, P. A., Lloyd, T., Geller, A., Raphael, S., & Glaser, J. (2016). *The science of justice: Race, arrests, and police use of force*. Center for Policing Equity.

Greenwald, A.G., Oakes, M.A., & Hoffman, H.G. (2003) Targets of discrimination: Effects of race on responses to weapons holders. *Journal of Experimental Social Psychology, 39,* 399-405.

Henry, P. J., & Sears, D. O. (2002). The symbolic racism 2000 scale. *Political Psychology, 23,* 253-283.

Hermans, D., Craske, M. G., Mineka, S., & Lovibond, P. F. (2006). Extinction in human fear conditioning. *Biological Psychiatry, 60,* 361-368.

Hermans, D., Vansteenwegen, D., & Eelen, P. (1999). Eye movement registration as a continuous index of attention deployment: Data from a group of spider anxious students. *Cognition & Emotion, 13,* 419–434.

Hofmann, S. G. (2008). Cognitive processes during fear acquisition and extinction in animals and humans: Implications for exposure therapy of anxiety disorders. *Clinical Psychology Review, 28,* 199-210.

Hugenberg, K., & Bodenhausen, G. V. (2003). Facing prejudice: Implicit prejudice and the perception of facial threat. *Psychological Science, 14,* 640-643.

Hugenberg, K. (2005). Social categorization and the perception of facial affect: target race moderates the response latency advantage for happy faces. *Emotion, 5,* 267-276.

Judd, C. M., Blair, I. V., & Chapleau K. M., (2004). Automatic stereotypes vs. automatic prejudice: Sorting out the possibilities in the Payne (2001) weapon paradigm. Journal of Experimental Social Psychology, 40, 75-81.

Knight, D.C., Waters, N., & Bandettini, P.A. (2009). Neural substrates of explicit and implicit fear memory. *Neuroimage, 45,* 208-214.

Krueger, J. (1996). Personal beliefs and cultural stereotypes about racial characteristics. *Journal of Personality and Social Psychology, 71,* 536-548.

Kveraga, K., Boshyan, J., Adams, R.B., Mote, J., Betz, N., Ward, N., ... & Barrett, L.F. (2015). If it bleeds, it leads: Separating threat from mere negativity. *Social Cognitive and Affective Neuroscience, 10,* 28-35.

Lai, C. K., Skinner, A. L., Cooley, E., Murrar, S., Brauer, M., Devos, T., ... & Nosek, B. A. (2016). Reducing implicit racial preferences: II. Intervention effectiveness across time. *Journal of Experimental Psychology: General, 145,* 1001.Rendon et al 2015

LeDoux, J. E. (1996). *The emotional brain: The mysterious underpinnings of emotional life*. New York, NY: Simon and Schuster.

LeDoux, J. E. (2012). Rethinking the emotional brain. *Neuron, 73*, 653-676.

LeDoux, J. E. (2014). Coming to terms with fear. *Proceedings of the National Academy of Sciences, 111*, 2871–2878.

Levin, B. (2021). Report to the nation: Anti-Asian prejudice and hate. Center fort he Study of Hate and Extremism. California State University San Barnardino. https://www.csusb.edu/sites/default/files/Report%20to%20the%20Nation%20-%20Anti-Asian%20Hate%202020%20Final%20Draft%20-%20As%20of%20Apr%2030%202021%206%20PM%20corrected.pdf

Löw, A., Weymar, M., & Hamm, A. O. (2015). When threat is near, get out of here dynamics of defensive behavior during freezing and active avoidance. *Psychological Science, 26,* 1706–1716.

Ma, D. S., Correll, J., & Wittenbrink, B. (2015). The Chicago face database: A free stimulus set of faces and norming data. *Behavior research methods, 47,* 1122-1135.

March, D. S., & Gaertner, L. (2021). A method for estimating the time of initiating correct categorization in mouse-tracking. *Behavior Research Methods. https://doi.org/10.3758/s13428-021-01575-9*

March, D. S., Gaertner, L., & Olson, M. A. (2017). In harm's way: On preferential response to threatening stimuli. *Personality and Social Psychology Bulletin, 43,* 1519-1529.

March, D. S., Gaertner, L., & Olson, M. A. (2018a). On the prioritized processing of threat in a Dual Implicit Process Model of evaluation. *Psychological Inquiry, 19,* 1-13.

March, D. S., Gaertner, L., & Olson, M. A. (2018b). Clarifying the explanatory scope of the Dual Implicit Process Model. *Psychological Inquiry, 29,* 38-43.

March, D., S., Olson, M., A., & Gaertner, L. (2020). Lions, and tigers, and implicit measures, Oh My! Implicit assessment and the valence vs. threat distinction. *Social Cognition, 38,* 154-164.

Méndez-Bértolo, C., Moratti, S., Toledano, R., Lopez-Sosa, F., Martínez-Alvarez, R., Mah, Y. H., ... & Strange, B. A. (2016). A fast pathway for fear in human amygdala. *Nature neuroscience, 19,* 1041.

Navarrete, C.D., Olsson, A., Ho, A.K., Mendes, W.B., Thomsen, L., & Sidanius, J. (2009). Fear extinction to an out-group face: The role of target gender. *Psychological Science, 20,* 155-158.

Neuberg, S.L., & Schaller, M. (2016). An evolutionary threat-management approach to prejudices. *Current Opinion in Psychology, 7*, 1-5.

O'Brien, F., & Cousineau, D. (2014) Representing error bars in within-subject designs in typical software packages. *The Quantitative Methods for Psychology, 10*, 58–70.

Öhman, A., & Mineka, S. (2001). Fears, phobias, and preparedness: toward an evolved module of fear and fear learning. *Psychological Review, 108*, 483-522.

Öhman, A., & Mineka, S. (2003). The malicious serpent: Snakes as a prototypical stimulus for an evolved module of fear. *Current Directions in Psychological Science, 12*, 5-9.

Olson, M. A., & Zabel, K. L. (2016). Measures of prejudice. In T. Nelson (Ed.), *Handbook of Prejudice, Stereotyping, and Discrimination* (pp. 175-212). New York: Psychology Press.

Payne, B. K. (2001). Prejudice and perception: The role of automatic and controlled processes in misperceiving a weapon. *Journal of Personality and Social Psychology, 81*, 181.

Plant, E. A., & Devine, P. G. (1998). Internal and external motivation to respond without prejudice. *Journal of personality and social psychology, 75*, 811-832.

Pettigrew, T. F., & Tropp, L. R. (2000). Does intergroup contact reduce prejudice? Recent meta-analytic findings. *Reducing prejudice and discrimination*, 93-114.

Phelps, E.P., O'Connor, K.J., Cunningham, W.A., Funayama, E. S., Gatenby, J.C., Gore, J.,C., & Banaji, M.R. (2000). Performance on indirect measures of race evaluation predicts amygdala activation. *Journal of Cognitive Neuroscience, 12*, 729-738.

Richeson, J.A., Todd, A.R., Trawalter, S., & Baird, A.A. (2008) Eye-gaze direction modulates race-related amygdala activity. *Group Processes and Intergroup Relations, 11*, 233-246.

Rozin, P. (1986). One-trial acquired likes and dislikes in humans: Disgust as a US, food predominance, and negative learning predominance. *Learning and Motivation, 17,* 180-189.

Ruscher, J. B. (2001). *Prejudiced communication: A social psychological perspective*. Guilford Press.

Sadler, M.S., Correll, J., Park, B., Judd, C.M. (2012). The world is not black and white: Racial bias in the decision to shoot in a multiethnic context. *Journal of Social Issues, 68,* 286-313.

Soderberg, C.K., Hennes, E.P., Lane, S.P. (2018). Sample size determination in psychological research. A workshop presented at the Society for the Improvement of Psychological Science. https://osf.io/ez2rm/

Thiem, K.C., Neel, R., Simpson, A.J., & Todd, A.R. (2019). Are Black women and girls associated with Danger? Implicit racial bias at the intersection of target age and gender. *Personality and Social Psychology Bulletin, 45,* 1427-1439.

Todd, A. R., Simpson, A.J., Thiem, K.C., & Neel, R. (2016). The generalization of implicit racial bias to young black boys: Automatic stereotyping or automatic prejudice? *Social Cognition, 34,* 306-323.

Todd, A. R., Thiem, K.C., & Neel, R. (2016). Does seeing faces of young Black boys facilitate the identification of threatening stimuli? *Psychological Science, 27,* 384-393.

Trawalter, S., Todd, A., Baird, A., & Richeson, J. (2008). Attending to threat: Race-based patterns of selective attention. *Journal of Experimental Social Psychology, 5,* 1322-1327.

Tukey, J. W. (1977). *Exploratory data analysis*. Reading: Addison-Wesley.

Unkelbach, C., Fiedler, K., Bayer, M., Stegmiller, M. & Danner, D. (2008). Why positive information is processed faster: The density hypothesis. *Journal of Personality and Social Psychology, 95*, 36-49.

Vuilleumier, P. (2005). How brains beware: neural mechanisms of emotional attention. *Trends in cognitive sciences, 9*, 585-594.

Westfall, J., Kenny, D.A., & Judd, C.M. (2014). Statistical power and optimal design in experiments in which samples of participants respond to samples of stimuli. Journal of Experimental Psychology: General, 143, 2020-2045.

Wittenbrink, B., Judd, C. M., & Park, B. (2001). Evaluative versus conceptual judgments in automatic stereotyping and prejudice. *Journal of experimental social psychology, 37*, 244-252.

Footnotes

1. A 2x3 multi-level ANOVA (with random intercept and target-slope for participants and random intercept for target-stimuli) yielded the same statistical conclusions and direction of effects for both studies (models would not converge with a random prime-slope for participants or intercept for prime-stimuli).

2. A sensitivity analysis for 80% power with 211 observations indicates that the smallest detectable effect size for the Black vs. White x Threat vs. Negative interaction is $f = .136$.

3. We "de-raced" new gun-images in two additional studies with photos we created (actor holding guns wearing long sleeves and gloves, and images cropped from chest to waist to obscure skin color). Both studies were identical to Studies 1 and 2; however, they swapped old for new gun images and one study ($N = 112$) added Asian face primes and the other did not ($N = 193$). Both replicated the tendency for Black vs. White faces to (a) yield slower responses to positive targets (ds = 0.60 & 0.47) and (b) to not differ in response to negative targets (ds = -0.10 & -0.11), but neither found a Black vs. White effect on threatening targets (ds = -0.11 & -0.13). We ran the study with the new gun-images and Asian faces after current Study 1, then reran it without Asian faces to assess if those faces contributed to the null race effect on threat, and then ran current Study 2 to ensure that the Study 1 race effect on threat replicated. A paired-comparison study of old vs. new gun-images indicated that participants ($N = 49$) were 2.2 times more likely to select old than new images as being scarier, which likely explains the null race effect on threat in the studies using the new images.

4. The rise of anti-Asian hostility in the US following the Covid-19 outbreak (Levin, 2021) is noteworthy and something to consider when comparing future data to the current data

which preceded the Covid-19 outbreak. However, we would suggest that the rise in Asian hostility is based on a contagion threat, not a physical safety threat (Neuberg & Schaller, 2016).

5. Due to a coding error on one computer, 48 participants experienced 30 (not 60) trials in the Cheerful/Depressed block (and for each race had 5, not 10, trials of happy-face / Cheerful target / Depressed distractor and sad-face / Depressed-target / Cheerful-distractor) bringing their total trials to 330 not 360. Two other participants (reason unknown) had 359 and 358 trials, respectively. This yields 41,037 trials across 118 participants.

6. $d_i = \sqrt{(x_i - x_f)^2 + (y_i - y_f)^2}$, where d_i is Euclidean distance at a given time interval, x_i and y_i is the horizontal and vertical location at a given time interval, and x_f and y_f is the horizontal and vertical location at the final time interval (i.e., location when the participant clicked in the target-label or, for distance from distractor, the corresponding location in the distractor-label).

7. The supplemental document reports the results of other mouse-tracking metrics (total response time, maximum deviation time, area under the curve, and maximum deviation). Unlike TICC (λ), those metrics do not reveal the point in time when participants began to categorize the faces uniquely in regard to the target label (i.e., time when participants began moving relatively closer to the target than distractor label; March & Gaertner, 2021).

8.The operative effect-size recommended for linear mixed effect models is d_0 (Westfall, Kenny, & Judd, 2014).

9. To appreciate the importance of purging the length of a name from the race of a name, we repeated the primary analysis with the exclusion of name-length as a covariate. Although the interaction remains significant, $F(1, 194) = 17.76$, $p = .0001$, $d_0 = -0.13$, the magnitude of the Black vs. White effect changes on both threatening and negative targets. Because the race-effect

absorbs the name-length effect, response times are slowed approximately 3 ms with Black primes, which have more letters, and sped approximately 3 ms with White primes, which have fewer letters. Consequently, the tendency for Black (vs. White) names to (a) facilitate responses to threatening targets is reduced by 6 ms and becomes non-significant $F(1, 23153) = 2.75$, $p = .0973$, $d_0 = -0.03$, ($M_{black} = 624$ ms vs. $M_{white} = 629$ ms) and (b) delay responses to negative targets is increased by 6 ms and becomes significant $F(1, 23153) = 23.41$, $p = .0001$, $d_0 = 0.09$, ($M_{black} = 671$ ms vs. $M_{white} = 661$ ms). Hence, confounding race with name-length yields the faulty conclusion that White Americans mores strongly associate White than Black men with negativity but not threat.

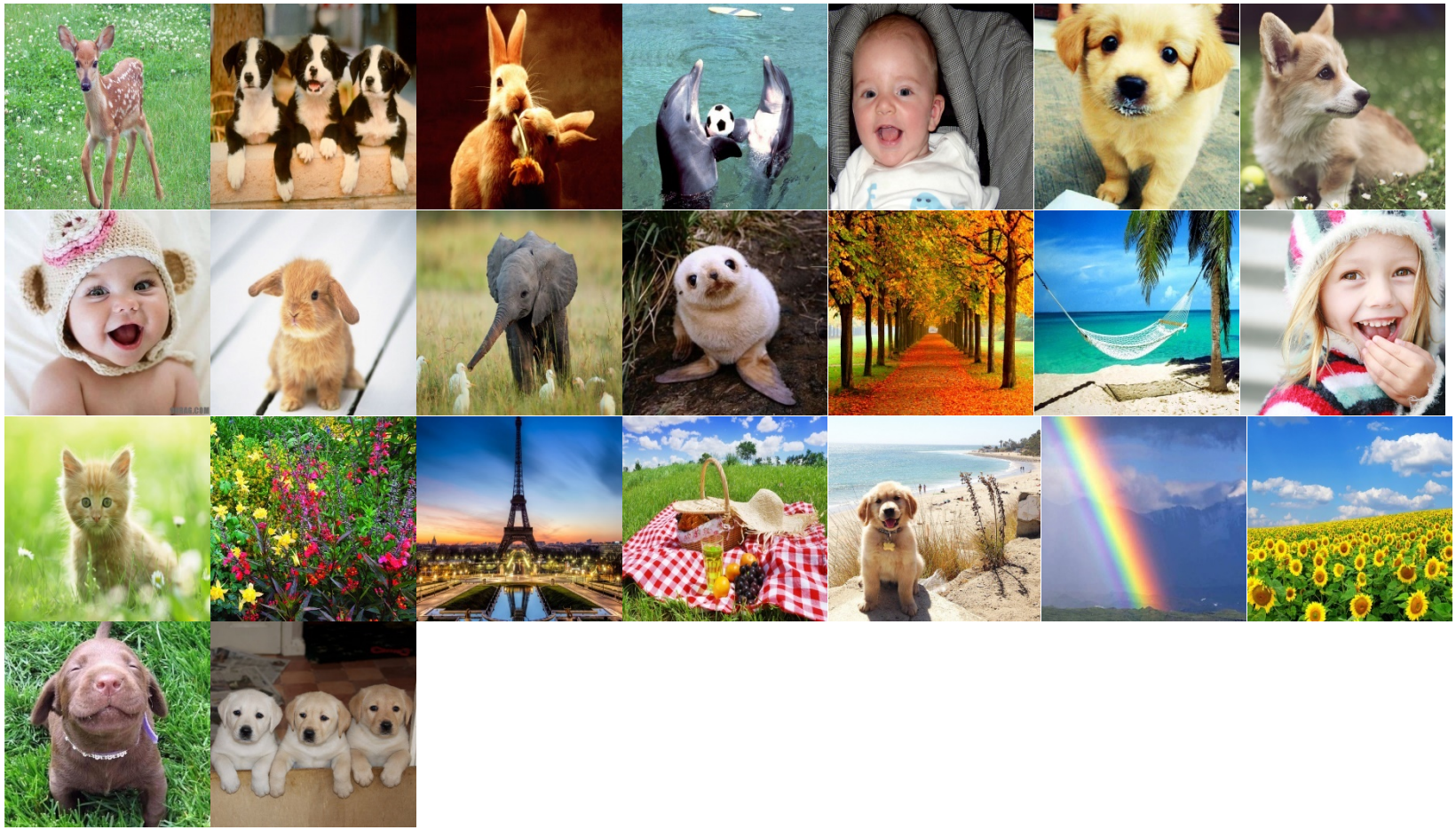
SUPPLEMENTAL MATERIALS

Stimuli used in Studies 1 and 2

Threatening Stimuli.

Negative Stimuli.

Positive Stimuli.



Facial Stimuli.

Black Faces.

White Faces.

Creating and Pilot Testing the Stimuli used in Study 3

No face databases could be located that contained the necessary stimuli, so we first created them. We gathered 20 neutral faces of each race (Asian, Black, White) from the Chicago Face Database (Ma et al., 2015). To create four expression categories (angry, happy, neutral, and sad) for each face, we created templates within FaceGen (a face morphing program) corresponding with emotional expressions described by the Facial Action Coding System (Ekman & Friesen, 1978). We applied each template to every neutral face to ensure that a given expression displayed roughly equal intensity across the faces (e.g., all angry faces were equally angry). This process involved several steps: (1) One at a time, each neutral face was imported into FaceGen and then overlaid over a fungible 3D template head. (2) The just-imported neutral face was first exported to ensure that it matched the look of the emotionally morphed faces (in terms of digitalization). (3) Each emotion template was applied to the neutral face, at which point (4) each newly morphed emotional face was exported. This process resulted in 60 faces of each expression (80 Asian, 80 Black, and 80 White; 240 total). These images were subsequently uniformly cropped to 450 x 650-pixels.

One-hundred and sixty-five subjects provided ratings of how angry, happy, or sad each face looked. We excluded 5 participants who responded faster than 500ms on >20% of their trials, resulting in 160 participants providing 38,053 ratings. We deleted individual ratings that were faster than 500ms (n = 666, 1.75%) or slower than 10000ms (n = 279, .73%), resulting in 37,108 usable ratings. Based on visual examination, we excluded 12 models (each of their 4 faces; 2 Asian, 3 Black, & 7 White) due to face-

morphing that caused them to appear abnormal (e.g., double nose, teeth bared, severe eye occlusion). We calculated a mean score of each rating for each face such that every face had 3 mean ratings. We then created Z-scores for each face within its emotional expression X rating type grouping using that group's mean and standard deviation. For example, separate Z-scores were created within the anger images for each rating of anger, happiness, and sadness, rendering three Z-scores for each face. We subsequently deleted 6 models (2 Asian, 1 White, and 3 Black) whose mean rating of any one of their three emotional faces fell >2SD above/below the mean of any of their three group means. We then excluded three models (3 Asian) whose neutral faces were rated greater than 2SD above the neutral group mean on any rating, leaving 14 Asian, 14 Black, and 12 White models. Lastly, we visually excluded models until each race group contained the 10 models used in the subsequently described study (see below for all images).

Angry Face Stimuli.

Sad Face Stimuli.

Happy Face Stimuli.

Neutral Face Stimuli.

Study 3 alternative Mouse-Tracking Metrics

The mouse tracking software provides for each trial metrics that describe the overall speed and shape of response. Specifically, each trial has an associated (1) reaction time (RT)[1], which is the duration from clicking the start button to clicking a response, (2) area under the curve (AUC), which is the total area between a hypothetical straight line running from the start-button to response and the actual path taken from start to response, and (3) the maximum deviation (MD) and maximum deviation time (MDT)[2], which quantifies the maximum distance from the hypothetical straight line to the actual response path and the time at which that occurred.

Like the main analyses, each target label was paired with one of the other three labels as distractors, target and distractor were not fully crossed (e.g., dangerous was paired with cheerful, depressed, and calm, but not itself). Consequently, we submitted each metric for each target label to separate 3(Race: Asian, Black, White) x 3(Distractor: three of Calm, Cheerful, Dangerous, Depressed) repeated measures ANOVAs.[3] We

[1] RT captures the amount of time it takes to complete a trial. As a start-to-finish metric, RT is likely influenced by several other characteristics of the response (e.g., velocity, attempts at late control). Such contamination renders the RT difficult to interpret. Lag time captures early reflexive movement while lessening the influence of other response characteristics, and therefore is a more appropriate metric for current research.

[2] Due to the dynamics of mouse-movement within a trial, the MD- and lag-times will often occur at a different time. For example, consider a trial where response movement stays close to the hypothetical straight-line response for much of the trial, but deviates from the straight line close to the response label. Or a trial where response mildly deviates toward the distractor label relative to the straight line but overcorrects to a degree that the MD occurs on the target side of the straight line. In such trials, the lag time will capture the initial turn toward the correct response, while the MD-time will capture the much later occurring MD. Lag captures initial turn toward the target, while MD captures the largest deviation from the hypothetical straight-line. Importantly, in many cases these will not occur at the same time.

[3] As with the main analyses, we also submitted each metric for each target label to separate 3(race) x 3(distractor-label) multi-level ANOVAs. For each analysis, we used Kenwood-Rodgers degrees of freedom and log-likelihood tests to determine which random effects, beyond a random intercept, were necessary. All models included a random intercept and random effect for distractor. In no instance did patterns of results from the multi-level analyses diverge from the reported repeated-measures approach.

present analyses of each metric separately within each separate target expression. Means for all Target x Distractor pairings can be found in Table S1.

Dangerous-target (angry face)

Reaction time. Reaction time for an angry face was influenced by a main effect of Race, $F(2, 107) = 11.82$, $p < .0001$, that was not moderated by the distractor type (i.e., Race x Distractor), $F(4, 105) = 1.02$, $p = .4023$. The Race main effect indicates that angry Black faces (M = 971ms) led to quicker identification than did White (M = 987ms), $F(1, 108) = 7.31$, $p = .0080$, $d = .41$, or Asian faces (M = 998ms), $F(1, 108) = 23.80$, $p < .0001$, $d = .57$, and quicker identification to White than Asian faces, $F(1, 108) = 3.42$, $p = .0670$, $d = .15$. The lack of an interaction indicates that reaction times for angry Black than White or Asian faces did not vary as a function of the distractor.

Maximum deviation time. Maximum deviation time for an angry face was influenced by a main effect of Race, $F(2, 107) = 12.76$, $p < .0001$, that was not moderated by the distractor type (i.e., Race x Distractor), $F(4, 105) = 0.48$, $p = .7538$. The Race main effect indicates that angry Black faces (M = 486ms) had an earlier maximum deviation time than did White (M = 497ms), $F(1, 108) = 8.60$, $p = .0041$, $d = .15$, or Asian faces (M = 505ms), $F(1, 108) = 25.51$, $p < .0001$, $d = .33$, and an earlier maximum deviation time to White than Asian faces, $F(1, 108) = 3.61$, $p = .0601$, $d = .21$. The lack of an interaction indicates that maximum deviation times for angry Black than White or Asian faces did not vary as a function of the distractor.

Area under the curve. Area under the curve for an angry face was influenced by a main effect of Race, $F(2, 107) = 29.87$, $p < .0001$, that was not moderated by the

distractor type (i.e., Race x Distractor), $F(4, 105) = .60$, $p = .6632$. The Race main effect indicates that angry Black faces ($M = .913$) led to smaller total area of divergence than did White ($M = 1.14$), $F(1, 108) = 39.96$, $p < .0001$, $d = .42$, or Asian faces ($M = 1.16$), $F(1, 108) = 52.24$, $p < .0001$, $d = .57$, and the latter two did not differ, $F(1, 108) = 1.85$, $p = .1763$, $d = .15$. The lack of an interaction indicates that area under the curve for angry Black than White or Asian faces did not vary as a function of the distractor.

Maximum deviation. Maximum deviation for an angry face was influenced by a main effect of Race, $F(2, 107) = 27.59$, $p < .0001$, that was not moderated by the distractor type (i.e., Race x Distractor), $F(4, 105) = .50$, $p = .7379$. The Race main effect indicates that angry Black faces ($M = .506$) led to smaller deviation from the optimal path than did White ($M = .581$), $F(1,108) = 38.07$, $p < .0001$, $d = .62$, or Asian faces ($M = .591$), $F(1, 108) = 46.69$, $p < .0001$, $d = .67$ and the latter two did not differ, $F(1, 108) = .94$, $p = .3335$, $d = .07$. The lack of an interaction indicates that maximum deviation for angry Black than White or Asian faces did not vary as a function of the distractor.

Depressed-target (sad face)

Reaction time. Reaction time for a sad face was influenced by a Race x Distractor interaction, $F(4, 105) = 5.07$, $p = .0009$. The Dangerous distractor led to *slower* responses when the sad face was Black ($M = 1032$ms) than when it was Asian ($M = 1011$ms), $F(1, 108) = 5.19$, $p = .0247$, $d = .24$. The Dangerous distractor did not differentially affect the reaction time to sad Black versus White ($M = 1021$ms), $F(1, 108) = 1.43$, $p = .2351$, $d = .13$, or Asian versus White faces, $F(1, 108) = .95$, $p = .3328$, $d = .08$. The Cheerful distractor led to *quicker* responses when the sad face was Black ($M =$

924ms) than when it was White (*M* = 953ms), *F*(1, 108) = 13.68, *p* =.0003, *d* = .37, or Asian (*M* = 949ms), *F*(1, 108) = 8.30, *p* =.0048, *d* = .31, and the latter two did not differ, *F*(1, 108) = 1.05, *p* =.3086, *d* = .11. The Calm distractor did not differentially affect the reaction time to identifying sad faces as a function of Race, *F*'s < 2.01, *p*'s > .1595, *d*'s < .17.

Maximum deviation time. Maximum deviation time for a sad face was influenced by a Race x Distractor interaction, *F*(4, 105) = 7.22, *p* < .0001. The Dangerous distractor led to a later maximum deviation time when the sad face was Black (*M* = 527ms) than when it was Asian (*M* = 505ms), *F*(1, 108) = 12.05, *p* =.0007, *d* =.35, and led to later maximum deviation time when the sad face was White (*M* = 518ms) than when it was Asian, *F*(1, 108) = 3.98, *p* =.0487, *d* = .15. The Dangerous distractor did not differentially affect the maximum deviation time to sad Black versus White faces, *F*(1, 108) = 1.90, *p* =.1706, *d* = .16. The Cheerful distractor led to earlier maximum deviation time when the sad face was Black (*M* = 458ms) than when it was White (*M* = 482ms), *F*(1, 108) = 15.48, *p* =.0001, *d* = .39, or Asian (*M* = 471ms), *F*(1, 108) = 5.91, *p* =.0167, *d* = .24, and the latter two did not differ, *F*(1, 108) = 2.87, *p* =.0933, *d* = .17. The Calm distractor did not differentially affect the maximum deviation time to identifying sad faces as a function of Race, *F*'s < 2.84, *p*'s > .0947 *d*'s < .17.

Area under the curve. Area under the curve for a sad face was influenced by a Race x Distractor interaction, *F*(4, 105) = 12.50 *p* <.0001. The Dangerous distractor led to a *larger* total area of deviance when the sad face was Black (*M* = 1.27) than when it was White (*M* = 1.05), *F*(1, 108) = 13.95, *p* =.0003, *d* = .29, or Asian (*M* = .998), *F*(1,

108) = 20.14, p <.0001, d = .38, and the latter two did not differ, F(1, 108) = .52, p =.4727, d = .09. The Cheerful distractor led to a *smaller* total area of deviance when the sad face was Black (M = .760) than when it was White (M = 1.02), F(1, 108) = 18.69, p <.0001, d = .43, or Asian (M = .975), F(1, 108) = 16.11, p <.0001, d = .38, and the latter two did not differ, F(1, 108) = .41, p =.5252, d = .08. The Calm distractor led to a smaller area of total deviance when the sad face was Black (M = 1.15) than when it was White (M = 1.294), F(1, 108) = 6.86, p =.0101, d = .16, or Asian (M = 1.295), F(1, 108) = 4.91, p =.0287, d = .18, and the latter two did not differ, F(1, 108) = .00, p =.9871, d = .06.

Maximum deviation. Maximum deviation for a sad face was influenced by a Race x Distractor interaction, F(4, 105) = 14.27 p <.0001. The Dangerous distractors led to a larger maximum deviation from the optimal path when the sad face was Black (M = .639) than when it was White (M = .553), F(1, 108) = 16.79, p <.0001, d = .34, or Asian (M = .547), F(1, 108) = 20.78, p <.0001, d = . 41, and the latter two did not differ, F(1, 108) = .07, p =.7928, d = .05. The Cheerful distractor led to smaller maximum deviation from the optimal when the sad face was Black (M = .442) than when it was White (M = .533), F(1, 108) = -19.57, p <.0001, d = .44, or Asian (M = .534), F(1, 108) = 20.01, p <.0001, d = .43, and the latter two did not differ, F(1, 108) = .01, p =.9213, d = .01. The Calm distractor led to smaller maximum deviation from the optimal path when the sad face was Black (M = .588) than when it was White (M = .643), F(1, 108) = 8.79, p =.0037, d = .22, or Asian (M =.640), F(1, 108) = 5.41, p =.0219, d =.20, and the latter two did not differ, F(1, 108) = .02, p = .8832, d = .03.

Cheerful-target (happy face)

Reaction time. Reaction for a happy face was influenced by a Race main effect, $F(2, 108) = 12.07$, $p < .0001$, that was not moderated by the distractor type (i.e., Race x Distractor), $F(4, 106) = 1.64$ $p = .1694$. Happy Black faces (M = 986ms) led to longer reaction times than did White (M = 959ms), $F(1, 109) = 24.20$, $p < .0001$, $d = .48$, or Asian faces (M = 969ms), $F(1, 109) = 9.32$, $p = .0029$, $d = .25$, and the latter two did not differ, $F(1, 109) = 3.14$, $p = .0792$, $d = .20$. The lack of an interaction indicates that reaction times for happy Black than White or Asian faces did not vary as a function of the distractor.

Maximum deviation time. Maximum deviation time for a happy face was influenced by a Race x Distractor interaction, $F(4, 106) = 5.69$, $p = .0003$. The Dangerous distractor led to later maximum deviation time when the happy face was Black (M = 503ms) than when it was White (M = 464ms), $F(1, 109) = 43.49$, $p < .0001$, $d = .57$, or Asian (M = 477ms), $F(1, 109) = 25.29$, $p < .0001$, $d = .40$, and later maximum deviation to Asian than White, $F(1, 109) = 6.91$, $p = .0090$, $d = .23$. The Depressed distractor led to later maximum deviation when the happy face was Black (M = 492ms) than when it was White (M = 479ms), $F(1, 109) = 4.38$, $p = .0387$, $d = .22$. The depressed distractor did not differentially affect maximum deviation time to differ to Black vs. Asian (M = 485ms), $F(1, 109) = 1.17$, $p = .2821$, $d = .10$, or Asian vs. White faces, $F(1, 109) = 1.37$, $p = .2248$, $d = .07$. The Calm distractor did not differentially affect response deviation to happy faces as a function of Race, Fs < 3.34, ps > .0703, d's < .20.

Area under the curve. Area under the curve for a happy face was influenced by a Race x Distractor interaction, $F(4, 106) = 3.21$, $p = .0157$. The Dangerous distractor led to

larger total area of divergence when the happy face was Black (M = 1.21) than when it was White (M = .809), $F(1, 109) = 39.81$, p <.0001, d = .36, or Asian (M = .976), $F(1, 109) = 14.47$, p =.0002, d = .36, and larger to Asian than White, $F(1, 109) = 10.27$, p =.0018, d = .15. The Depressed distractor led to a larger total area of divergence when the happy face was Black (M = 1.19) than when it was White (M = .938), $F(1, 109) = 9.18$, p =.0031, d = .35, or Asian (M = .962), $F(1, 109) = 13.71$, p =.0003, d = .28, and the latter two did not differ, $F(1, 109) = .11$, p =.7459, d = .03. The Calm distractor did not differentially affect the area under the curve to happy faces as a function of Race, ts < 1.90, ps > .1712, ds < .11.

Maximum deviation. Maximum deviation for a happy face was influenced by a Race x Distractor interaction, $F(4, 106) = 3.30$, p =.0137. The Dangerous distractor led to larger deviation when the happy face was Black (M = .627) than when it was White (M = .469), $F(1, 109) = 44.60$, p <.0001, d = .46, or Asian (M = .530), $F(1, 109) = 17.17$, p <.0001, d = .40, and larger deviation to Asian than White, $F(1, 109) = 9.50$, p =.0026, d = .16. The Depressed distractor led to larger deviation when the happy face was Black (M = .614) than when it was White (M = .497), $F(1, 109) = 22.43$, p <.0001, d = .45, or Asian (M = .517), $F(1, 109) = 11.62$, p = .0009, d = .32, and the latter two did not differ, $F(1, 109) = .55$, p =.4600, d =.07. The Calm distractor did not differentially affect response deviation to happy faces as a function of Race, Fs < 2.70, ps > .1031, d's < .14.

Calm-target (neutral face)

Reaction time. Reaction time for a neutral face was influenced by a Race x Distractor interaction, $F(2, 106) = 6.36$, p =.0001. The Dangerous distractor led to slower

responses when the neutral face was Black (*M* = 1009ms) than when it was White (*M* = 981ms), *F*(1, 109) = 5.48, *p* =.0210 *d* = .21, or Asian (*M* = 993ms), *F*(1, 109) = 3.17, *p* =.0780, *d* = .15, and the latter two did not differ, *F*(1, 109) = .89, *p* =.3477, *d* = .09. The Depressed distractor led to slower responses when the neutral face was Black (*M* = 1018ms) then when it was Asian (*M* = 1040ms), *F*(1, 109) = 7.20, *p* =.0084, *d* = .26. The Depressed distractor did not differentially affect response times to Black versus White (*M* = 1021ms) faces, *F*(1, 109) = .08, *p* =.7763, *d* = .04, or White versus Asian faces, *F*(1, 109 = 4.66, *p* =.0331, *d* = .22. The Cheerful distractor led to quicker responses when the neutral face was Black (*M* = 1003ms) than when it was Asian (*M* = 981ms), *F*(1, 109) = 6.46, *p* =.0124, *d* = .26. The Cheerful distractor did not differentially affect reaction times to Black versus White (*M* = 1005ms), *F*(1, 109) = 8.94, *p* =.0034, *d* = .32, or White versus Asian faces, *F*(1, 109) = .02, *p* =.8906, *d* = .05.

Maximum deviation time. Maximum deviation time for a neutral face was influenced by a Race x Distractor interaction, *F*(4, 106) = 6.16, *p* =.0002. The Dangerous distractor led to later maximum deviation when the neutral face was Black (*M* = 522ms) than when it was White (*M* = 502ms), *F*(1, 109) = 13.45, *p* =.0004, *d* = .35, or Asian (*M* = 502ms), *F*(1, 109) = 12.19, *p* =.0007, *d* = .31, and the latter two did not differ, *F*(1, 109) = 0.01 *p* =.9208, *d* =.04. The Cheerful distractor led to later maximum deviation when the neutral face was Black (*M* = 503ms) than when it was Asian (*M* = 492ms), *F*(1, 109) = 3.78, *p* =.0544, *d* = .19, and earlier when it was Asian than White (*M* = 513ms), *F*(1, 109) = 16.55, *p* <.0001, *d* = .42. The Cheerful distractor did not differentially affect maximum deviation time to Black vs. White, *F*(1, 109) = 2.67, *p* =.1048, *d* = .14. The

Depressed distractor did not differentially affect response deviation to happy faces as a function of Race, *F*s < 2.70, *p*s > .1031, *d*s < .19.

Area under the curve. Area under the curve for a neutral face was influenced by a Race x Distractor interaction, $F(4,106) = 4.69$ $p = .0016$. The Dangerous distractor led to a larger total area of deviance when the neutral face was Black ($M = 1.28$) than when it was White ($M = .984$), $F(1,109) = 29.08$, $p < .0001$, $d = .51$, or Asian ($M = 1.04$), $F(1,109) = 20.75$, $p < .0001$, $d = .46$, and the latter two did not differ, $F(1,109) = 1.60$, $p = .2089$, $d = .10$. The Depressed distractor led to a larger total area of deviance when the face was Black ($M = 1.22$) than when it was White ($M = 1.08$), $F(1,109) = 6.50$, $p = .0122$, $d = .25$, and area of divergence did not differ to either Black versus Asian, $F(1,109) = 2.88$, $p = .0925$, $d = .14$, or Asian vs White faces, $F(1,109) = .72$, $p = .3979$, $d = .10$. The Cheerful distractor did not lead to differential areas under the curve as a function of Race, *t*s < 1.04, *p*s > .3098, *d*s < .17.

Maximum deviation. Maximum deviation for a neutral face was influenced by a Race x Distractor interaction, $F(4, 106) = 5.49$, $p = .0005$. The Dangerous distractor led to larger deviation when the neutral face was Black ($M = .637$) than when it was White ($M = .520$), $F(1, 109) = 34.16$, $p < .0001$, $d = .55$, or Asian ($M = .542$), $F(1, 109) = 22.50$, $p < .0001$, $d = .50$, and the latter two did not differ, $F(1, 109) = 1.36$, $p = .2461$, $d = .09$. The Depressed distractor led to larger deviation when the neutral face was Black ($M = .6361$) than when it was White ($M = .569$), $F(1, 109) = 9.05$, $p = .0033$, $d = .30$, or Asian ($M = .694$), $F(1, 109) = 3.00$, $p = .0859$, $d = .15$, and the latter two did not differ, $F(1, 109) =$

1.65, p =.2013, d = .14. The Cheerful distractor did not differentially affect response deviation to happy faces as a function of Race, Fs < .95, ps > .3315, ds < .14.

Table S1. Study 3 Mean mouse-tracking metrics for each Target x Distractor pairing as a function of race.

Target (and Facial Expression)	**Metric**	**Distractor**											
		Dangerous			Depressed			Cheerful			Calm		
		Asian	Black	White	Asian	Black	White	Asian	Black	White	Asian	Black	White
Dangerous (Angry)	TICC				573	515	544	510	483	503	530	495	524
	RT				1040	1002	1016	965	945	964	990	965	982
	MD Time				530	508	517	485	470	483	498	479	492
	AUC				1.23	0.990	1.21	1.07	0.862	1.07	1.18	0.889	1.07
	MD				.615	.530	.606	.559	.488	.568	.599	.499	.567
Depressed (Sad)	TICC	522	576	521				498	467	512	573	555	558
	RT	1011	1032	1021				949	924	959	1038	1038	1026
	MD Time	505	527	518				471	458	482	526	530	519
	AUC	0.998	1.27	1.04				0.975	0.760	1.02	1.29	1.15	1.29
	MD	.547	.639	.553				.534	.442	.533	.640	.588	.643
Cheerful (Happy)	TICC	492	548	481	500	548	500				530	521	509
	RT	950	982	946	964	986	952				993	991	980
	MD Time	477	503	464	485	492	479				502	495	492
	AUC	0.976	1.21	0.809	0.963	1.19	0.938				1.03	1.06	.981
	MD	.530	.627	.469	.517	.614	.497				.540	.555	.519
Calm (Neutral)	TICC	523	567	518	551	573	529	519	522	531			
	RT	993	1009	985	1040	1018	1021	981	1003	1005			
	MD Time	502	522	502	526	527	518	492	503	513			
	AUC	1.01	1.24	.950	1.12	1.22	1.08	0.996	1.01	1.04			
	MD	.542	.637	.520	.594	.631	.569	.539	.541	.558			

Note. RT = reaction time in milliseconds; MDT = maximum deviation time; AUC = area under the curve; MD = maximum deviation.

Study 4 alternative Mouse-Tracking Metrics

Like the main analyses, we separately entered participants' RT, MD-time, MD, and AUC estimates for a given block in into a 2(Race: Black vs. White) x 2(Within-block Face: positive vs. neutral; angry vs. neutral; sad vs. neutral; angry vs. sad) repeated-measures ANOVA (degrees of freedom vary due to missing λ estimates from non-convergence or missing trajectories). Table S2 displays the average of each metric for each race of face in each block.

Positive / Not-Positive (Happy and Neutral Faces) Block

Reaction time. Reaction time to the target label was not influenced by a Race x Face interaction, $F(1, 320) = 1.87$, $p = .1729$. Instead, reaction time was affected by only two main effects. Regardless of the race of the face, participants were to quicker to classify as Positive happy faces (M = 955ms) than Not-Positive for neutral faces (M = 1000ms), $F(1, 320) = 142.43$, $p < .0001$, $d = .67$. Regardless of the emotion of the face, participants were quicker to classify Black (M = 974ms) than White faces (M = 980ms), $F(1, 320) = 3.42$, $p = .0653$, $d = .10$.

Maximum deviation time. Maximum deviation time to the target label was not influenced by a Race x Face interaction, $F(1, 320) = 0.23$, $p = .6311$. Instead, maximum deviation time was affected by only one main effect. Regardless of the race of the face, the maximum deviation occurred earlier to Positive happy faces (M = 468ms) than Not-Positive for neutral faces (M = 496ms), $F(1, 320) = 129.42$, $p < .0001$, $d = .64$.

Area under the curve. Area under the curve to the target label was influenced by a Race x Face interaction, $F(1, 320) = 7.01$, $p = .0085$. Total area under the curve was

larger to Positive when the happy face was Black (*M* = .885) than White (*M* = .837), *F*(1, 320) = 2.61, *p* =.1073, *d* = .09, and was smaller to Not-Positive when the neutral face was Black (*M* = 1.025) than White (*M* = 1.087), *F*(1, 320) = 4.49, *p* =.0349, *d* = .12.

Maximum deviation. Maximum deviation to the target label was influenced by a Race x Face interaction, *F*(1, 320) = 9.34, *p* =.0024. Maximum deviation was larger to Positive when the happy face was Black (*M* = .489) than White (*M* = .466), *F*(1, 320) = 3.70, *p* =.0553, *d* = .11, and was smaller to Not-Positive when the neutral face was Black (*M* = .554) than White (*M* = .581), *F*(1, 320) = 6.18, *p* =.0135, *d* = .13.

Dangerous / Not-Dangerous (Angry and Neutral Faces) Block

Reaction time. Reaction time to the target label was not influenced by a Race x Face interaction, *F*(1, 319) = 0.02, *p* =.8753. Instead, reaction time was affected by only one main effect. Regardless of the race of the face, reaction time was quicker to Dangerous angry faces (*M* = 984ms) than Not-Dangerous for neutral faces (*M* = 1015ms), *F*(1, 319) = 52.48, *p* <.0001, *d* = .41.

Maximum deviation time. Maximum deviation time to the target label was influenced by a Race x Face interaction, *F*(1, 319) = 9.22, *p* =.0026. Maximum deviation time was earlier to Dangerous when the angry face was Black (*M* = 486ms) than White (*M* = 496ms), *F*(1, 319) = 6.90, *p* =.0090, *d* = .14, and was later to Not-Dangerous when the neutral face was Black (*M* = 508ms) than White (*M* = 502ms), *F*(1, 319) = 2.54, *p* =.1121, *d* = .09.

Area under the curve. Area under the curve to the target label was not influenced by a Race x Face interaction, *F*(1, 319) = 0.22, *p* =.6412. Instead, area under

the curve was affected by only one main effect. Regardless of the race of the face, the area under the curve was smaller to Dangerous angry faces (M = 1.01) than Not-Dangerous for neutral faces (M = 1.07), $F(1, 319) = 5.78$, p =.0168, d = .13.

Maximum deviation. Maximum deviation to the target label was not influenced by a Race x Face interaction, $F(1, 319) = 0.56$, p =.4538. Instead, maximum deviation was affected by only one main effect. Regardless of the race of the face, the maximum deviation was smaller to Dangerous for angry faces (M = .534) than Not-Dangerous for neutral faces (M = .572), $F(1, 319) = 13.33$, p =.0003, d = .20.

Negative / Not-Negative (Sad and Neutral Faces) Block

Reaction time. Reaction time to the target label was influenced by a Race x Face interaction, $F(1, 309) = 9.41$, p =.0024. Reaction time was quicker to Negative when the sad face was White (M = 1019ms) than Black (M = 1031ms), $F(1, 309) = 4.22$, p =.0407, d = .11, and was slower to Not-Negative when the neutral face was White (M = 1059ms) than Black (M = 1047ms), $F(1, 309) = 5.02$, p =.0257, d = .13.

Maximum deviation time. Maximum deviation time to the target label was influenced by a Race x Face interaction, $F(1, 309) = 5.52$, p =.0194. Maximum deviation time occurred earlier to Negative when the sad face was White (M = 514ms) than Black (M = 521ms), $F(1, 309) = 3.62$, p =.0579, d = .09, and was later to Not-Negative when the neutral face was White (M = 532ms) than Black (M = 527ms), $F(1, 309) = 1.81$, p =.1789, d = .09.

Area under the curve. Area under the curve to the target label was not influenced by a Race x Face interaction, $F(1, 309) = 1.17$, $p = .2795$, nor where there any main effects.

Maximum deviation. Maximum deviation to the target label was not influenced by a Race x Face interaction, $F(1, 309) = 1.37$ $p = .2422$. Instead, maximum deviation was affected by only one main effect. Regardless of the race of the face, the maximum deviation was smaller to Negative for sad faces ($M = .572$) than Not-Negative for neutral faces ($M = .598$), $F(1, 309) = 5.38$, $p = .0210$, $d = .09$.

Dangerous / Negative (Angry and Sad Faces) Block

Reaction time. Reaction time to the target label was not influenced by a Race x Face interaction, $F(1, 316) = 1.72$, $p = .1907$. Instead, reaction time was affected by only one main effect. Regardless of the race of the face, reaction time was quicker to Dangerous angry faces ($M = 981$ms) than Negative for sad faces ($M = 1002$ms), $F(1, 316) = 25.38$, $p < .0001$, $d = .29$.

Maximum deviation time. Maximum deviation time to the target label was influenced by a Race x Face interaction, $F(1, 316) = 10.03$, $p = .0017$. Maximum deviation time occurred earlier to Dangerous when the angry face was Black ($M = 490$ms) than White ($M = 499$ms), $F(1, 316) = 4.76$, $p = .0298$, $d = .13$, and was later to Negative when the sad face was Black ($M = 507$ms) than White ($M = 498$ms), $F(1, 316) = 5.48$, $p = .0199$, $d = .12$.

Area under the curve. Area under the curve to the target label was influenced by a Race x Face interaction, $F(1, 316) = 28.24$, $p < .0001$. Area under the curve was smaller

to Dangerous when the angry face was Black (M = .973) than White (M = 1.12), F(1, 316) = 21.97, p <.0001, d = .24, and was larger to Negative when the sad face was Black (M = 1.12) than White (M = 1.01), F(1, 316) = 10.78, p =.0011, d = .17.

Maximum deviation. Maximum deviation to the target label was influenced by a Race x Face interaction, F(1, 316) = 33.66, p <.0001. Maximum deviation was smaller to Dangerous when the angry face was Black (M = .531) than White (M = .589), F(1, 316) = 25.84, p <.0001, d = .29, and was larger to Negative when the sad face was Black (M = .589) than White (M = .544), F(1, 316) = 13.19, p =.0003, d = .21.

Table S2. Study 4 Mean mouse-tracking metrics for each face in each block as a function of race.

	Dangerous/Not-Dangerous				Negative/Not-Negative			
	Angry Face		Neutral Face		Sad Face		Neutral Face	
Metric	Black	White	Black	White	Black	White	Black	White
TICC	496	518	538	525	520	535	554	565
RT	981	988	1012	1017	1031	1019	1047	1059
MDT	486	496	508	502	521	514	527	532
AUC	.995	1.02	1.07	1.07	1.12	1.12	1.12	1.16
MD	.527	.541	.572	.573	.575	.569	.592	.605

	Positive/Not-Positive				Dangerous/Negative			
	Happy Face		Neutral Face		Angry Face		Sad Face	
Metric	Black	White	Black	White	Black	White	Black	White
TICC	485	465	512	533	496	530	541	522
RT	954	955	994	1005	977	1004	984	1000
MDT	468	469	495	498	490	499	507	498
AUC	.885	.837	1.03	1.09	.973	1.12	1.12	1.01
MD	.489	.467	.554	.581	.531	.589	.589	.544

Note. RT = reaction time in milliseconds; MDT = maximum deviation time; AUC = area under the curve; MD = maximum deviation

Pilot Testing the Stimuli used in Study 5

Name Piloting. Ninety-three White subjects participated for partial credit in a psychology course. Seated in computer cubicles, they rated 20 Black and 20 White names (Leavitt & Dubner; 2014) on how typical each name is of being Black and/or White. Participants were told, "We are gathering names for use in future studies. The purpose of the current pilot study is to gather ratings on how people think different names are typical of either being Black and/or White. We will use these ratings to build categories for these names. You will be shown several names. Your task is to rate how typical each name of being either Black and/or White." They were subsequently shown each name in a random order and rated how Black and/or White each was on separate 0 (Not Typical) to 6 (Very Typical) scales. We averaged the Black and White rating for each name, respectively. We retained the six Black names with the highest average Black rating and six White names with the highest average White rating for use in Study 5. See Table S3.

Table S3. Average Black and White ratings for each name used in Study 5.

Name	Black Rating	White Rating
Jamal	5.27	0.66
DeShawn	5.24	0.63
DeAndre	5.22	0.77
Tyrone	4.84	0.96
Trevon	4.81	1.16
Darnell	4.73	1.10
Jack	1.81	5.12
Connor	1.60	5.13
Jake	1.72	5.19
Ethan	1.77	5.19
Scott	1.54	5.24
Brad	1.11	5.26

Target Word Piloting. Twenty-two White subjects rated 67 words on how much each signified Danger and/or Negativity. Participants were told, "We are gathering words for use in future studies. The purpose of the current pilot study is to gather ratings on how people think different words are associated with the concepts of negativity and/or danger. We will use these ratings to build categories for these words. You will be shown several words. Your task is to rate how much each word differently signifies Negativity and/or Danger." They were subsequently shown each word in a random order and rated how much each word signified negativity and/or danger on separate 0 (Not at all) to 6 (Very much) scales. We averaged the Negativity and Dangerous rating for each word, respectively, and based on those ratings, assigned each word to a category. Negative category words had negative ratings greater than 4 and dangerous ratings less than 3. Dangerous category words had negative ratings greater than 4 and dangerous ratings greater than 4. From those that qualified for each category, we chose 6 negative and 6 dangerous words for use in Study 5. See Table S4.

Table S4. Average Negative and Dangerous ratings for each target word used in Study 5.

Word	Negativity Rating	Dangerous Rating
Undesirable	5.14	1.62
Displeasing	5.14	1.76
Lousy	5.00	1.24
Inferior	5.00	1.86
Awful	5.67	2.81
Disliked	4.95	1.62
Aggressive	4.90	5.28
Violent	5.48	5.71
Threatening	5.28	5.48
Murderous	6.00	6.00
Harmful	5.38	5.52
Unsafe	4.71	5.81