Original Manuscript

Social Psychological and
Personality Science
1-7

DOI: 10.1177/1948550619900574
journals.sagepub.com/home/spp
SAGE

Tracking Prejudice: A Mouse-Tracking Measure of Evaluative Conflict Predicts Discriminatory Behavior

David E. Melnikoff[1], Thomas C. Mann[2], Paul E. Stillman[3], Xi Shen[4], and Melissa J. Ferguson[4]

Abstract

Explicit evaluations of racial out-groups often involve conflict between opposing evaluative tendencies. Yet this type of conflict is difficult to capture with standard measures of evaluative processing, which either ignore explicit evaluation or capture only the aspects of explicit evaluation that are consciously accessible and freely reported. A new tool may fill this gap in our ability to measure conflict in racial evaluation. This tool, called the mouse-tracking measure of racial bias (Race-MT), is designed to capture conflict in explicit evaluations of racial groups, even if that conflict is neither consciously accessible nor freely reported. We vetted the Race-MT by exploring whether it predicts discriminatory behavior. Across five studies (four preregistered, $N = 1{,}492$), we used the Race-MT to measure conflict in people's positive, explicit evaluations of racial out-groups versus in-groups. These measures predicted discriminatory behavior in a noisy, naturalistic setting, suggesting that the Race-MT provides theoretically meaningful and predicatively useful insights into racial evaluation.

Keywords

discrimination, automatic/implicit processes, conflict, prejudice/stereotyping, attitudes

Ask someone if they like people from other racial groups, and they will likely say yes. Yet evaluations of racial out-groups often are far more complex than such reports would suggest. Usually they reflect mixtures of positive and negative feelings, mirroring both the complexity of social stimuli, and intrapersonal tension between egalitarian ideals and racial bias (e.g., Fiske et al., 2002). These feelings likely emerge over the course of evaluative processing before culminating in a final judgment. So even when people truthfully report that they like racial out-groups overall, their responses may obscure conflict between opposing evaluative tendencies—feelings of like and dislike that occur simultaneously or in quick succession. Such conflict may be both common and predictive of a range of discriminatory behaviors (Stillman et al., 2018), yet it is difficult to capture with the field's most popular measurement tools: self-report and indirect attitude measures.

Self-report measures directly ask people about their feelings toward racial groups. In principle, this approach could be used to measure conflict in racial evaluations, but it has several drawbacks. First, self-report measures require that people be consciously aware of the degree of conflict in their racial evaluations, a requirement that may be satisfied sometimes but not always (Hahn & Gawronski, 2019; Hahn et al., 2014). Second, self-report measures rely on people to willingly express violations of egalitarian ideals and thus are susceptible to self-presentational pressures (Banaji & Greenwald, 2016; Fazio et al., 1995). Among people who disregard such pressures (Kteily & Bruneau, 2017), self-reports may reliably estimate conflict in racial evaluations, but among many other people, especially in cultures with strong egalitarian norms, self-reports may be misleading.

Unlike self-reports, indirect attitude measures specifically avoid asking people about their feelings toward racial groups. They tap such feelings indirectly by measuring how racial stimuli (e.g., pictures of people from racial out-groups) influence people's responding to nonracial stimuli (e.g., positive and negative words). This approach minimizes the influence of self-presentation and makes no assumptions about conscious awareness, solving some of the problems with self-report. But, like all measurement tools, indirect attitude measures have their limitations.

[1] Department of Psychology, Yale University, New Haven, CT, USA
[2] Department of Psychology, Harvard University, Cambridge, MA, USA
[3] School of Management, Yale University, New Haven, CT, USA
[4] Department of Psychology, Cornell University, Ithaca, NY, USA

Corresponding Author:
David E. Melnikoff, Department of Psychology, Yale University, New Haven, CT 06511, USA.
Email: david.melnikoff@yale.edu

One limitation of indirect attitude measures is that they do not capture conflict that unfolds while people explicitly evaluate racial groups. Indeed, indirect attitude measures are specifically designed to ensure that the intention to evaluate racial groups is absent (or minimized) throughout the procedure. This approach is indispensable as it is uniquely capable of capturing an important form of evaluative processing. At the same time, there are downsides to ignoring conflict that unfolds during explicit racial evaluation. For instance, without measuring this sort of conflict, it may be difficult to predict behavior in situations where people explicitly evaluate members of racial out-groups. Consistent with this, recent studies have questioned whether indirect attitude measures reliably predict discrimination, particularly in noisy, "real-world" environments, where people are free to explicitly evaluate others (Oswald et al., 2013, 2015; cf. Kurdi et al., 2018).

Due to the limitations of self-report and indirect attitude measures, we may be missing an important aspect of conflict in racial evaluation. In particular, neither approach measures conflict that people are unwilling or unable to report, and which emerges during explicit evaluative processing. How might this sort of conflict be captured?

One promising solution is a new measure of racial bias that leverages a technique called mouse-tracking (MT; see Freeman, 2018; Freeman & Ambady, 2009, 2010; Freeman et al., 2016; Hehman et al., 2014, 2015; Spivey & Dale, 2006; Stillman et al., 2018). Here is how it works: Participants explicitly evaluate racial groups, and the trajectory of their computer mouse is recorded as they move it toward a "like" button or a "dislike" button. Almost all participants report liking racial out-groups and in-groups. However, participants tend to move their mouse cursor closer to the dislike button while evaluating racial out-groups, revealing a unique form of racial bias: greater conflict in the unfolding of positive, explicit evaluations of racial out-groups relative to in-groups (Wojnowicz et al., 2009). We call this tool the mouse-tracking measure of racial bias (hereafter, the Race-MT).

By requiring people to explicitly evaluate others and then covertly measuring the generation of their responses, the Race-MT may capture conflict during explicit racial evaluation (like self-reports) without requiring that the conflict be consciously accessible or willingly disclosed (like indirect attitude measures). In this article, we test the predictive potential of the Race-MT in order to establish it as a new measure of racial evaluative processes. Although recent research in other domains has provided initial evidence that mouse movements can predict behavior (Hehman et al., 2015; Stillman & Ferguson, 2019; Stillman et al., 2017), it is unknown if this extends to the racial domain, raising the question of whether the Race-MT captures conflict in racial evaluation that is theoretically meaningful and predictively useful. We addressed this question by conducting the first-ever tests of the Race-MT's ability to predict subsequent acts of discrimination.

Methods and Results

We conducted five studies and preregistered four. All studies followed a similar procedure. First, we recruited White participants online via Amazon Mechanical Turk. Second, we used the Race-MT to measure pro-White/anti-Black bias, operationalized as the degree of conflict in participants' positive, explicit evaluations of Black people relative to White people. Third, we introduced participants to a White or Black target person and observed whether or not participants helped them. In what follows, we describe each of these steps in detail. We then describe the unique elements of each individual study plus its results. Finally, we present the results of an integrative data analysis.

Measuring Racial Bias

We measured racial bias using the Race-MT developed by Wojnowicz and colleagues (2009). This task involved a series of evaluation decisions. Each decision included two response options—Dislike and Like—which were located at the top-left and top-right corners of an 800×600 px response window, respectively. To begin each trial, participants clicked on a small box labeled "START," which was located at the bottom-center of the response window. After clicking, the box was replaced by a stimulus word or noun phrase, which participants were asked to evaluate by moving their mouse cursor to the Dislike or Like response option.

The task included 24 stimuli, 2 of which were target stimuli. One target stimulus was the name of participants' own racial group ("White people") and the other was the name of a racial out-group ("Black people"). There were 11 positively valenced distractors (e.g., "Freedom," "Love"), 5 of which were normatively positive social groups (e.g., "Asian people," "Jewish people"). The final 11 stimuli were negatively valenced distractors (e.g., "War," "Poison"), 5 of which were normatively negative social groups (e.g., "Nazis," "Terrorists"). All stimuli were presented multiple times with the number of repetitions varying across studies (see below for details). Specifically, each stimulus was presented once across multiple blocks (within each block, each stimulus was selected without replacement).

We computed the degree of conflict during each response in terms of *maximum deviation* (MD).[1] To compute MD, we recorded the *x*- and *y*-coordinates of participants' mouse cursor movements during each evaluation decision. We then prepared these data in line with standard practices (Freeman & Ambady, 2010). Specifically, we time-normalized the trajectories into 101 time bins and rescaled every response, such that each trajectory terminated at the top-right response location. MD is calculated as the largest perpendicular deviation, out of all time steps, between the actual trajectory and an idealized straight line between the start and endpoints.

Next, we removed responses that met any of the following preregistered criteria: (i) over 500 ms until initial mouse movement, (ii) over 2 s to make a response, (iii) 3 or more standard deviations (*SD*s) above or below the mean MD (or area under the curve [AUC], see footnote 1) on that trial type, (iv) mouse cursor moved outside of the 800 × 600 px response area, or (v) "incorrect" (i.e., Like responses to normatively negative stimuli and Dislike responses to normatively positive stimuli including White people and Black people). Twenty-three percent of responses to target stimuli were excluded across all five studies. Similar MT-based tasks that have been conducted in laboratory settings have had lower exclusion rates, suggesting that such tasks generate significantly more errors and outliers when run online.

After making the above exclusions, we converted MD into an estimate of racial bias: MDb. Specifically, we subtracted each participant's mean score on White people trials from their mean score on Black people trials. Positive scores denote greater conflict during positive, explicit evaluations of Black people versus positive, explicit evaluations of White people and thus correspond to pro-White/anti-Black bias. Conceivably, effects of MDb on behavior could be driven solely by MD on White people trials, which would run counter to our prediction that conflict during evaluations of out-groups predicts discriminatory behavior. We ruled out this possibility with additional analyses that did not rely on difference scores (see Supplemental Analyses). These analyses confirmed that the effects of MDb were driven by conflict on both types of trials.

A potential criticism of the Race-MT is that conflict during explicit evaluation can be measured just as accurately, and more easily, simply by measuring how long it takes people to report their judgments. To explore this possibility, we recorded not only MD but also reaction time (RT), operationalized as the time (in ms) that elapsed between pressing the START button and selecting a response option. We converted RT into an estimate of racial bias—RTb—by subtracting each participant's mean score on White people trials from their mean score on Black people trials.

Manipulating Race

After completing the Race-MT, participants received the following message:

> You finished the main task sooner than the average time, so if you wouldn't mind, I have included some simple transcription tasks in this hit, and you are welcome to do as many or as few as you would like. These transcriptions are for research being done by a graduate student in my lab. His introduction and instruction are on the following page.

On the following screen, participants saw a message from a (fake) graduate student. Included in the message was the student's name, which we manipulated to be stereotypically Black (DeAndre) or White (Dustin). The purpose of manipulating only the student's name was to signal Target Race in a subtle and ecologically valid manner.

Measuring Helping

Participants received the following message:

> Hello! My name is [DeAndre/Dustin] and I am looking for some help with my dissertation. Specifically, I am trying to develop a computer program that can transcribe text from images. To assess how well my program works, I have to compare the quality of its transcriptions to the quality of transcriptions that humans produce by hand. Accordingly, I need people who are willing to help me by transcribing text from images. My dissertation depends on it!
>
> Please transcribe as much text as you like—the more you transcribe the better, but even a little would be a major help. Thank you!

Below the message was the following question, "Would you like to help me by transcribing some text?" Participants could select "Yes, I would like to help by transcribing some text," or "No, I would like to skip directly to the end of the survey." If participants indicated they wanted to help, they received a paragraph to transcribe.[2] Thus, our outcome measure was a single, binary choice to help or not help someone achieve an important goal—a choice that is both consequential and highly "noisy" due to the fact that it is not repeated and is multiply determined. In this way, our experiment was designed to assess relationships between racial bias and discriminatory behavior in a noisy, naturalistic setting.

What our single-item, binary-choice dependent measures adds in ecological validity, it subtracts in statistical power (Payne et al., 2016). To address this, we preregistered all but our first (pilot) study and base our conclusions on an integrative data analysis. We have not excluded any studies from the present report; every study that we have conducted in line with the above procedures is reported below and included in our integrative data analysis.

Study 1

Study 1 followed all of the procedures described above, with the stimuli in the Race-MT presented 4 times each. We recruited 451 participants, 229 of whom met our inclusion criteria of being Caucasian and having used a computer mouse to complete the survey. Thirty-six participants provided no usable data (i.e., they selected Dislike on all White people trials and/or all Black people trials), resulting in a final sample of 193 (age: $M = 38$, $SD = 12$; sex: 45% female).

Replicating past findings (Wojnowicz et al., 2009), we found that MDb ($M = .29$, 95% confidence interval [CI] [.24, .34], $t(192) = 10.91$, $p < .001$) was significantly greater than zero, meaning that partcipants showed greater curvature toward Dislike on Black people trials versus White people trials. We also found that RTb ($M = 156$, 95% CI [130, 182], $t(192) = 11.77$, $p < .001$) was significantly greater than zero, meaning that participants took longer to select Like on Black people trials versus White people trials. Finally, we found that RTb was correlated but not redundant with MDb, $r(191) = .502$, $p < .001$. This means that the more participants curved toward

Dislike on Black people trials relative to White people trials, the longer participants took to select Like on Black people trials relative to White people trials.

To assess whether MDb predicts behavior, we regressed MDb, Target Race, and their interaction term on a binary outcome variable denoting whether or not participants chose to help. We (separately) ran an equivalent model to assess whether RTb predicts discrimination. If MDb or RTb predicts discriminatory behavior, then its interaction with Target Race should be significant. This interaction may involve a positive association with the likelihood of helping the White target (if the discrimination entails selectively helping in-group members) and/or a negative association with the likelihood of helping the Black target (if the discrimination entails selectively withholding help from out-group members).

MDb predicted discriminatory behavior in Study 1, but RTb did not. Specifically, we found a significant MDb × Target Race interaction ($b = .69$, $SE = .32$, $p = .029$, odds ratio [OR] = 2.0) and a nonsignificant RTb × Target Race interaction ($b = .18$, $SE = .33$, $p = .583$, $OR = 1.20$). MDb had a more negative relationship with helping behavior toward the Black target ($b = -.32$, $SE = .25$, $p = .186$, $OR = .72$) versus the White target ($b = .37$, $SE = .20$, $p = .066$, $OR = 1.45$). Although the RTb × Target Race interaction was not significant, the results were in the predicted direction: RTb had a more negative relationship with helping behavior toward the Black target ($b = -.29$, $SE = .24$, $p = .230$, $OR = 0.75$) versus the White target ($b = -.12$, $SE = .21$, $p = .592$, $OR = 0.89$).

Study 2

In Study 2, we explored whether we could replicate our findings from Study 1 using a shorter Race-MT, one in which all stimuli were presented twice rather than 4 times each. We also preregistered our recruitment plan, methods, analyses, and predictions (https://aspredicted.org/dq7av.pdf). We recruited 349 participants, 275 of whom met our inclusion criteria of being Caucasian and having used a computer mouse to complete the survey. Sixty-two participants provided no usable data (i.e., they selected Dislike on every "White person" trial and/or on every "Black person" trial), resulting in a final sample of 213 (age: $M = 38$, $SD = 12$; sex: 55% female).

As in Study 1, we found that MDb ($M = .27$, 95% CI [.20, .33], $t(213) = 8.23$, $p < .001$) and RTb ($M = 149$, 95% CI [116, 182], $t(213) = 8.92$, $p < .001$) were significantly greater than zero. We also found that RTb was strongly correlated but not redundant with MDb, $r(212) = .583$, $p < .001$. Additionally, neither MDb nor RTb predicted discriminatory behavior. Both the Target Race × MDb interaction ($b = .12$, $SE = .22$, $p = .604$, $OR = 1.12$) and the Target Race × RTb interaction ($b = .08$, $SE = .23$, $p = .742$, $OR = 1.08$) were nonsignificant. The results were, however, in the predicted direction. MDb had a less positive relationship with helping behavior toward the Black target ($b = .11$, $SE = .15$, $p = .454$, $OR = 1.12$) versus the White target ($b = .23$, $SE = .16$, $p = .163$, $OR = 1.26$). RTb also had a less positive relationship with helping behavior toward the Black target ($b = .13$, $SE = .15$, $p = .396$, $OR = 1.14$) versus the White target ($b = .21$, $SE = .18$, $p = .241$, $OR = 1.23$).

Study 3

In Study 3, we modified the Race-MT to be briefer than the one used in Study 1 but more reliable than the much shorter task we tried in Study 2. Specifically, we presented all distractor stimuli twice (rather than 4 times each, as in Study 1) and all target stimuli 4 times each (rather than twice, as in Study 2). We also preregistered our recruitment plan, methods, analyses, and predictions (https://aspredicted.org/k63ir.pdf). We recruited 504 participants, 472 of whom met our inclusion criteria of being Caucasian and having used a computer mouse to complete the survey. Seventy-four participants provided no usable data (i.e., every White person or Black person trial was excluded), resulting in a final sample of 398 (age: $M = 37$, $SD = 12$; sex: 49% female).

As in Studies 1 and 2, we found that MDb ($M = .24$, 95% CI [.21, .27], $t(397) = 15.79$, $p < .001$) and RTb ($M = 121$, 95% CI [106, 137], $t(397) = 15.24$, $p < .001$) were significantly greater than zero and that RTb was strongly correlated but not redundant with MDb, $r(396) = .467$, $p < .001$. Additionally, neither MDb nor RTb was significant the predictor of discriminatory behavior. Neither the Target Race × MDb interaction ($b = .30$, $SE = .25$, $p = .222$, $OR = 1.35$) nor the Target Race × RTb interaction ($b = -.20$, $SE = .25$, $p = .422$, $OR = 0.82$) was significant. The results for MDb, however, were in the predicted direction: MDb had a more negative relationship with helping behavior toward the Black target ($b = -.27$, $SE = .17$, $p = .107$, $OR = 0.76$) versus the White target ($b = .03$, $SE = .18$, $p = .859$, $OR = 1.03$). The results for RTb were not in the predicted direction: RTb had a less negative relationship with helping behavior toward the Black target ($b = -.004$, $SE = .17$, $p = .979$, $OR = 0.996$) versus the White target ($b = -.21$, $SE = .18$, $p = .265$, $OR = 0.81$).

In the next two studies (Studies 4 and 5), we attempted to replicate Studies 2 and 3, respectively, with a single modification: the addition of an attention check. Specifically, we included an item at the end of the survey to confirm that participants read the name of the person who solicited help. We asked participants to identify the person's name from a list of six names, three of which were stereotypically White (including Dustin) and three of which were stereotypically Black (including DeAndre). We decided a priori to exclude all participants who failed to recall the graduate student's name. As in Studies 2 and 3, we preregistered our recruitment plan, exclusion criteria, methods, analyses, and predictions (Study 4: https://aspredicted.org/8vw2y.pdf, Study 5: https://aspredicted.org/nk6hy.pdf).

Study 4

This study was a replication of Study 2, in which we used the shorter task with only two trials. By adding an attention check, we can examine whether that study did not show an effect because of the shorter task or because we failed to adequately exclude people who were not paying attention to the critical dependent variable. We recruited 499 participants, 457 of whom met our inclusion criteria of being Caucasian and having used a computer mouse to complete the survey. Eighty-eight participants provided no usable data (i.e., they selected Dislike on all White people trials and/or all Black people trials), and 122 participants failed the attention check, resulting in a final sample of 247 (age: $M = 37$, $SD = 11$; sex: 46% female). Due to the relatively high rate of incorrect responses on the attention check, we ran additional analyses and confirmed that (1) participants failed the attention at a similar rate across conditions ($b = -.01$, $SE = .20$, $p = .962$, $OR = 0.99$) and (2) all significant effects reported below remain significant when including participants who failed the attention check.

As in Studies 1–3, we found that MDb ($M = .29$, 95% CI [.24, .35], $t(247) = 10.51$, $p < .001$) and RTb ($M = 118$, 95% CI [88, 148], $t(247) = 7.72$, $p < .001$) were significantly greater than zero. We also found that RTb was strongly correlated but not redundant with MDb, $r(246) = .601$, $p < .001$.

Just like in Study 2, neither MDb nor RTb was the significant predictor of discriminatory behavior. Neither the Target Race × MDb interaction ($b = .35$, $SE = .22$, $p = .114$, $OR = 1.42$) nor the Target Race × RTb interaction ($b = .001$, $SE = .21$, $p = .994$, $OR = 1.001$) was significant. The results were, however, in the predicted direction. MDb had a more negative relationship with helping behavior directed toward the Black target ($b = -.13$, $SE = .16$, $p = .418$, $OR = 0.88$) versus the White target ($b = .22$, $SE = .15$, $p = .147$, $OR = 1.25$). The same pattern held for the relationship between RTb and helping: slightly more negative when the target person was Black ($b = -.07$, $SE = .15$, $p = .642$, $OR = 0.933$) versus White ($b = -.069$, $SE = .15$, $p = .643$, $OR = 0.934$). It could be that the shorter task is not sufficiently reliable to capture conflict that predicts behavior. Therefore, in the next study, we incorporated an attention check into a longer MT task in our replication of Study 3.

Study 5

Study 5 was a replication of Study 3 with the addition of the attention check described in the previous study (Study 4). Because the predicted effects had emerged directionally multiple times and significantly only once, we recruited a larger sample in Study 5 in order to increase our ability to detect a significant effect. Specifically, we recruited 800 participants, 747 of whom met our inclusion criteria of being Caucasian and having used a computer mouse to complete the survey. One hundred twenty participants provided no usable data (i.e., they selected Dislike on every White person trial or on every Black person trial), and 188 participants failed the attention check, resulting in a final sample of 439 (age: $M = 40$, $SD = 31$; sex: 52% female). Due to the relatively high rate of incorrect responses on the attention check, we ran addition analyses and confirmed that (i) participants failed the attention at a similar rate across conditions ($b = -.66$, $SE = .16$, $p = .676$, $OR = 0.936$) and (ii) all significant effects reported below remain significant when including participants who failed the attention check.

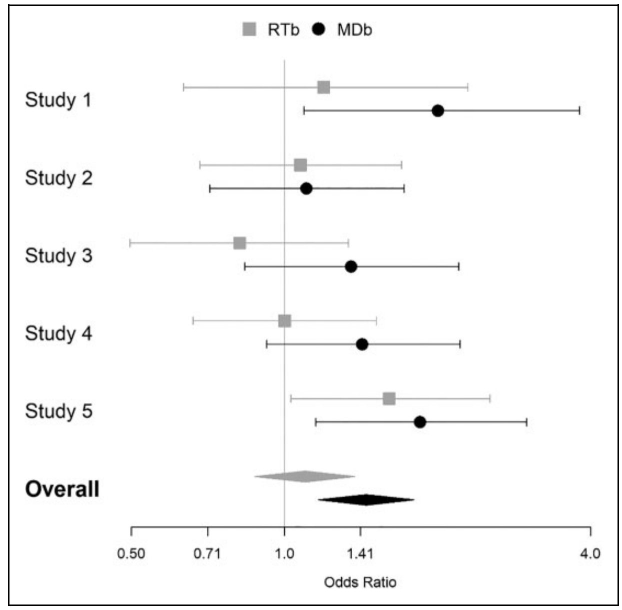

Figure 1. Odds ratios and 95% confidence intervals for Bias × Target Race interactions on decision to help as a function of study and measurement type (MDb vs. RTb).

As in Studies 1–4, we found that MDb ($M = .25$, 95% CI [.22, .28], $t(438) = 16.96$, $p < .001$) and RTb ($M = 132$, 95% CI [116, 149], $t(438) = 15.69$, $p < .001$) were significantly greater than zero and that RTb was correlated but not redundant with MDb, $r(437) = .456$, $p < .001$. Additionally, MDb and RTb predicted discriminatory behavior: Both the Target Race × MDb interaction ($b = .61$, $SE = .24$, $p = .012$, $OR = 1.85$) and the Target Race × RTb interaction ($b = .47$, $SE = .23$, $p = .039$, $OR = 1.61$) were significant. MDb had a more negative relationship with helping behavior toward the Black target ($b = -.13$, $SE = .17$, $p = .446$, $OR = 0.88$) versus the White target ($b = .48$, $SE = .17$, $p = .004$, $OR = 1.62$). RTb also had a more negative relationship with helping behavior toward the Black target ($b = -.14$, $SE = .16$, $p = .39$, $OR = 0.87$) versus the White target ($b = .33$, $SE = .16$, $p = .039$, $OR = 0.62$).

Integrative Data Analysis

To reliably estimate the ability of MDb and RTb to predict subsequent discriminatory behavior, we subjected the data from Studies 1–5 to an integrative data analysis. Specifically, we fit a generalized linear mixed effect model with a random intercept for study and random slopes for racial bias (MDb or RTb), Target Race, and their interaction term (Figure 1).

These analyses revealed that MDb predicted discriminatory behavior, but RTb did not. Specifically, we found a significant MDb × Target Race interaction ($b = .37$, $SE = .11$, $p = .0007$,

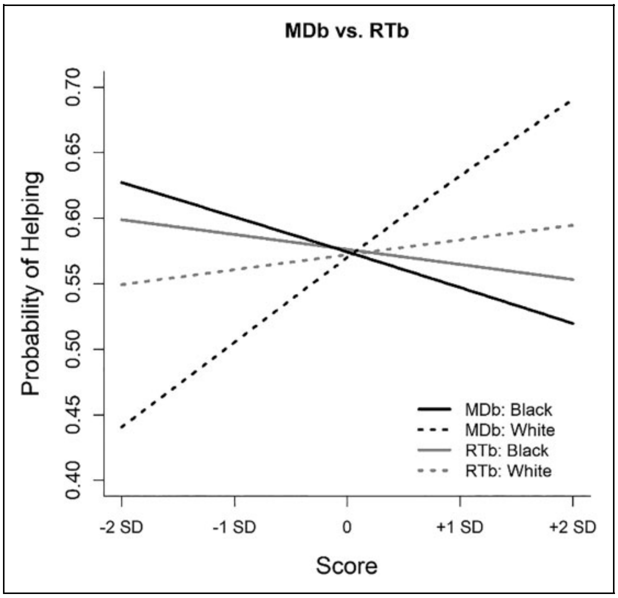

Figure 2. Relationships between (standardized) bias scores and probability of helping across Studies 1–5 as a function of metric (MDb vs. RTb) and Target Race (Black vs. White).

$OR = 1.45$), such that MDb had a more negative relationship with helping toward the Black target ($b = -.11$, $SE = .08$, $p = .158$, $OR = 0.90$) versus the White target ($b = .26$, $SE = .08$, $p = .002$, $OR = 1.29$). The pattern of discrimination was consistent with the selective allocation of help to the in-group (see Figure 2). The RTb × Target Race interaction was not significant ($b = .09$, $SE = .12$, $p = .416$, $OR = 1.10$; see Figure 2), although the pattern of results was in the predicted direction: RTb had a more negative relationship with helping toward the Black target ($b = -.05$, $SE = .07$, $p = .525$, $OR = 0.95$) versus the White target ($b = .02$, $SE = .11$, $p = .859$, $OR = 1.02$). Moreover, the MDb × Target Race interaction remained significant when adjusting for the nonsignificant RTb × Target Race interaction ($b = .45$, $SE = .14$, $p = .001$, $OR = 1.57$). Finally, the results of a paired samples t-test suggest that MDb is significantly better at predicting discriminatory behavior than RTb. Specifically, the *OR*s associated with the MDb × Target Race interactions across Studies 1–5 were on average larger than those associated with the RTb × Target Race interactions, $t(4) = 3.17$, $p = .034$.

Discussion

An integrative data analysis of five studies ($N = 1{,}492$) found that the Race-MT predicts discriminatory behavior in a noisy, naturalistic setting. These findings are the first to show that the Race-MT captures conflict in racial evaluation that is theoretically meaningful and predictively useful, establishing the Race-MT as an important new addition to the social psychologist's tool kit. We also found that the Race-MT outperformed a simpler, RT-based approach, suggesting that, in some circumstances, the Race-MT may be a more reliable measure of conflict in racial evaluation. Collectively, these findings suggest that the Race-MT is uniquely capable of capturing conflict that people are unable or unwilling to report, and which emerges during the explicit evaluation of racial groups. These results further add to the growing body of research pointing to the predictive validity of motor movements for real-world behaviors (see Freeman, 2018; Stillman et al., 2018).

Two caveats bear mentioning. First, the effect sizes relating the Race-MT to behavior were small. This limits the practical utility of the Race-MT and calls for future efforts to improve its reliability. Numerous adjustments may help, such as increasing the total number of critical trials or administering in a lab environment rather than online. Second, the ability of the Race-MT to predict behavior likely depends on numerous undiscovered moderators. For instance, the ability of the Race-MT to predict behavior may depend on the context in which that behavior occurs—in contexts where people do not explicitly evaluate others, the Race-MT may be outperformed by indirect attitude measures. This hypothesis illustrates the promise of Race-MT as a guide for future research. Having been validated, it can now be used to explore novel and important questions about the underlying dynamics of racial evaluation.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

David E. Melnikoff https://orcid.org/0000-0002-7090-9995

Supplemental Material

The supplemental material is available in the online version of the article.

Notes

1. We also measured conflict with a metric called area under the curve (AUC). AUC and maximum deviation yielded equivalent results and were almost perfectly correlated across all studies (all *r*s > .9). We report the results of AUC in Supplemental Analyses.
2. Of 865 participants who chose to help across our five studies, only 9 failed to transcribe any text, 49 transcribed less than half the text, and 813 transcribed all the text.

References

Banaji, M. R., & Greenwald, A. G. (2016). *Blindspot: Hidden biases of good people*. Bantam.

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.

Fiske, S. T., Cuddy, A. J., Glick, P., & Xu, J. (2002). A models of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. *Journal of Personality and Social Psychology*, *82*, 878–902.

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

Freeman, J. B., & Ambady, N. (2009). Motions of the hand expose the partial and parallel activation of stereotypes. *Psychological Science*, *20*(10), 1183–1188.

Freeman, J. B., & Ambady, N. (2010). MouseTracker: Software for studying real-time mental processing using a computer mouse-tracking method. *Behavior Research Methods*, *42*, 226–241.

Freeman, J. B., Pauker, K., & Sanchez, D. T. (2016). A perceptual pathway to bias: Interracial exposure reduces abrupt shifts in real-time race perception that predict mixed-race bias. *Psychological Science*, *27*, 502–517.

Hahn, A., & Gawronski, B. (2019). Facing one's implicit biases: From awareness to acknowledgment. *Journal of Personality and Social Psychology*, *116*, 769–794.

Hahn, A., Judd, C. M., Hirsh, H. K., & Blair, I. V. (2014). Awareness of implicit attitudes. *Journal of Experimental Psychology: General*, *143*, 1369–1392.

Hehman, E., Carpinella, C. M., Johnson, K. L., Leitner, J. B., & Freeman, J. B. (2014). Early processing of gendered facial cues predicts the electoral success of female politicians. *Social Psychological and Personality Science*, *5*, 815–824.

Hehman, E., Stolier, R. M., & Freeman, J. B. (2015). Advanced mouse-tracking analytic techniques for enhancing psychological science. *Group Processes & Intergroup Relations*, *18*, 384–401.

Kteily, N. S., & Bruneau, E. (2017). Darker demons of our nature: The need to (re)focus attention on blatant forms of dehumanization. *Current Directions in Psychological Science*, *26*, 487–494.

Kurdi, B., Seitchik, A. E., Axt, J., Carroll, T., Karapetyan, A., Kaushik, N., Tomezsko, D., Greenwald, A., & Banaji, M. R. (2018). Relationship between the Implicit Association Test and intergroup behavior: A meta-analysis. *The American Psychologist*, *74*, 569–586.

Oswald, F. L., Mitchell, G., Blanton, H., Jaccard, J., & Tetlock, P. E. (2013). Predicting ethnic and racial discrimination: A meta-analysis of IAT criterion studies. *Journal of Personality and Social Psychology*, *105*, 171–192.

Oswald, F. L., Mitchell, G., Blanton, H., Jaccard, J., & Tetlock, P. E. (2015). Using the IAT to predict ethnic and racial discrimination: Small effect sizes of unknown societal significance. *Journal of Personality and Social Psychology*, *108*, 562–571.

Payne, B. K., Brown-Iannuzzi, J. L., & Loersch, C. (2016). Replicable effects of primes on human behavior. *Journal of Experimental Psychology: General*, *145*, 1269–1279.

Spivey, M. J., & Dale, R. (2006). Continuous dynamics in real-time cognition. *Current Directions in Psychological Science*, *15*, 207–211.

Stillman, P. E., & Ferguson, M. J. (2019). Decisional conflict predicts impatience. *Journal of the Association for Consumer Research*, *4*(1), 47–56.

Stillman, P. E., Medvedev, D., & Ferguson, M. J. (2017). Resisting temptation: tracking how self-control conflicts are successfully resolved in real time. *Psychological science*, *28*, 1240–1258.

Stillman, P. E., Shen, X., & Ferguson, M. J. (2018). How mouse-tracking can advance social cognitive theory. *Trends in Cognitive Sciences*, *22*, 531–543.

Wojnowicz, M. T., Ferguson, M. J., Dale, R., & Spivey, M. J. (2009). The self-organization of explicit attitudes. *Psychological Science*, *20*, 1428–1435.

Author Biographies

David E. Melnikoff received his PhD in social psychology at Yale University. Currently, he is a postdoctoral researcher at Northeastern University. His work focuses on the computational principles of evaluative processing and goal-directed cognition.

Thomas C. Mann received his PhD in social psychology at Cornell University. He was most recently a National Science Foundation post-doctoral research fellow at Harvard University, and investgates implicit attitude change.

Paul E. Stillman is a Postdoctoral Associate at the Yale School of Management. His research seeks to promote optimal self-regulation by understanding the processes underlying how people navigate conflicting motives. His research takes an interdisciplinary approach, integrating neuroscience, implicit, behavioral, and network methodologies.

Xi Shen is a PhD student in social psychology at Cornell University. She examines implicit impression formation and updating toward different groups, and from different sources, with a focus on its cognitive underpinnings and behavioral consequences.

Melissa J. Ferguson is an experimental social psychologist. She received her doctorate in social psychology from New York University and then joined the psychology department at Cornell University. Her research focuses on the implicit cognitive processes that enable evaluation, goal-pursuit, self-control, and social behavior.

Handling Editor: Joseph Cesario