NOTICE: This article is the accepted version of a work that was published in Personality and Social Psychology Bulletin (PSPB). The reference to the published version is: Röhner, J., & Lai, C. K. (2021). A diffusion model approach for understanding the impact of 17 interventions on the race Implicit Association Test. *Personality and Social Psychology Bulletin, 47*(9), 1374-1389. https://doi.org/10.1177/0146167220974489
Supplements to this work can be found at: https://osf.io/9xegp/

A Diffusion Model Approach for Understanding the Impact of 17 Interventions on the Race Implicit Association Test

Jessica Röhner and Calvin K. Lai

University of Bamberg, Washington University in St. Louis

Author Note

Jessica Röhner, Department of Psychology, University of Bamberg; Calvin K. Lai, Department of Psychological & Brain Sciences, Washington University in St. Louis.

The authors would like to cordially thank Joel Le Forestier and Philipp J. Thoss for their help in analyzing the data.

Correspondence concerning this article should be addressed to Jessica Röhner, Department of Psychology, University of Bamberg, Markusplatz 3, 96047 Bamberg, Germany. E-Mail: jessica.roehner@uni-bamberg.de

Abstract

Performance on implicit measures reflects construct-specific and non-construct-specific processes. This creates an interpretive issue for understanding interventions to change implicit measures: change in performance could reflect changes in the constructs-of-interest or changes in other mental processes. We re-analyzed data from six studies (N = 23,342) to examine the process-level effects of 17 interventions and one sham intervention to change race Implicit Association Test (IAT) performance. Diffusion models decompose overall IAT performance (D-scores) into construct-specific (ease of decision-making), and non-construct-specific processes (speed-accuracy tradeoffs, non-decision-related processes like motor execution). Interventions that effectively reduced D-scores changed ease of decision-making on compatible and incompatible trials. They also eliminated differences in speed-accuracy tradeoffs between compatible and incompatible trials. Non-decision-related processes were impacted by two interventions only. There was little evidence that interventions had any long-term effects. These findings highlight the value of diffusion modeling for understanding the mechanisms by which interventions affect implicit measure performance.

Keywords: changes in overall IAT performance, diffusion model analyses, IAT_v, IAT_a, IAT_{t_0}

A Diffusion Model Approach for Understanding the Impact of 17 Interventions on the Race Implicit Association Test

People can explicitly report no racial preferences but nonetheless express racial preferences on implicit measures[1] (Devine, 1989; Fazio et al., 1995; Nosek et al., 2007). These implicit measures have predicted a variety of behavior, such as friendliness in interracial interactions (e.g., Dovidio et al., 2002), biases in medical decision-making (e.g., Green et al., 2007), and hiring discrimination (e.g., Rooth, 2010). Several meta-analyses examining hundreds of research reports have found robust and reliable evidence for relationships between implicit measures and behavior (Cameron et al., 2012; Greenwald et al., 2009; Forscher, Lai, et al., 2019; Kurdi et al., 2019; Oswald et al., 2013). This has sparked interest in understanding how to change performance on implicit measures (Lai et al., 2013).

Decades of research have documented changes on implicit measures in response to new experiences (Lai et al., 2013; Sritharan & Gawronski, 2010). To advance research, two large scale research projects (Lai et al., 2014, and Lai et al., 2016; referred to as L2014, and L2016 in the rest of this article) with a total of six studies compared 18 interventions (17 interventions and one faking intervention) against a control condition in their ability to change overall performance on race Implicit Association Tests (IATs). Overall IAT performance was assessed using the *D*-score algorithm, which examines relative differences in the speed of categorizing stimuli on the basis of race and valence. It measures association strength by computing the mean difference in reaction times between incompatible and compatible[2] IAT phases divided by their overall standard deviation (Greenwald et al., 2003a, 2003b). Of the 18 interventions, nine were effective at immediately changing *D*-scores toward neutrality.

[1] By *implicit measures*, we are referring to the large class of social cognition measures that assess associative constructs indirectly without requiring participants to actively bring to mind a target association (Forscher, Lai et al., 2019; Greenwald & Lai, 2020).

[2] On the IAT, compatible trials are the trials that are theoretically more in line with pre-existing biases (e.g., pairing White with Good and Black with Bad on the Race IAT) and the incompatible trials are the trials that are theoretically more in conflict with pre-existing biases (e.g., pairing White with Bad and Black with Good on the Race IAT).

Effective interventions tended to give *participants experiences with counterstereotypical people*, used *evaluative conditioning*, or gave participants *intentional strategies to override bias*. Ineffective interventions tended to give participants opportunity to *reflect on egalitarian values*, *engage with others' perspectives*, or *induced emotion*.

Effective interventions were tested again to see if their effects persisted after several days in two studies (L2016). To the researchers' surprise, none of the nine interventions that were effective at immediately reducing *D*-scores had durable effects that persisted beyond a couple of days. Together, findings from these large-scale investigations indicated that interventions were effective at changing *D*-scores temporarily but without long-lasting effects.

Diffusion Model Analyses

D-scores are summary scores that reflect a mixture of components, some of which are related to theoretical processes-of-interest and others of which are not (Calanchini & Sherman, 2013; Röhner et al., 2011). Thus, they reflect *whether or not* overall IAT performance was changed. They do not allow investigating *how* interventions changed IAT performance. To address this gap, researchers can employ diffusion models. Unlike *D*-scores which primarily rely on response times, diffusion models use both response times and response accuracy to assess the mechanisms underlying decisions in binary decision tasks such as IATs (e.g., Klauer et al., 2011), in which participants have to categorize stimuli (e.g., Black faces vs. White faces) to one of two categories (e.g., Black vs. White).

In diffusion models, the decision process includes actual decision-making and the processes that precede and follow decision-making (e.g., encoding of stimuli; motor execution of response). Decisions (e.g., Does the Black face belong to the category White or to the category Black?) are based on evidence collection in information sampling processes. Evidence refers to bits of information retrieved from the stimulus (e.g., Black face) that are interpreted in light of knowledge from memory as to whether they support one of the two categories (e.g., Black vs. White). In contrast to *D*-scores, diffusion model analyses allow for

the detailed inspection of decision-making processes (e.g., Röhner & Ewers, 2016b). Consequently, diffusion models have been applied to a variety of decision tasks such as recognition memory (e.g., Spaniol et al., 2006) and perceptual decision-making (e.g., Germar et al., 2014) to name only two here. For detailed introductions to diffusion modeling and tutorials, see e.g. Ratcliff (2014), Röhner and Ewers (2016a), Voss et al. (2013), or Wagenmakers et al. (2007). To understand decision-making, diffusion models are comprised of three main parts: v (ease of decision-making), a (response caution), and t_0 (non-decision-related processes), that are reflected in diffusion-model-based IAT effects (Table 1).

Parameter *v (ease of decision-making)*. This parameter measures the speed of information intake (Voss et al., 2010), which incorporates the ability to perform the decision and the ease of the task (Klauer et al., 2011). The higher the v, the faster the participant decided while simultaneously committing only few errors (i.e., the easier it was to make the decision). Ratcliff (1978) described parameter v as the only parameter that represents input from memory into the decision system. In other words, parameter v is related to the activation (and/or inhibition) of associative memory content that drive the ease of decision-making. If the activated associations match the task, then the task is easier. If the activated associations do not match the task, then the task is more difficult. For example, in race IATs automatic processes may bias perceptual analyses to focus on race early on, which affects the ease of the task. If a person sees a Black face, mental processes may then extract the meaning of the stimulus and co-activate ‘Black’ with ‘Bad’ (Klauer, 2014). On compatible trials, this biases the decision process and makes it easier to categorize Black with Bad. On incompatible trials, the difficulty of the task increases because cognitive processes drive decision processes in conflicting directions due to the co-activation of Black and Bad that pulls categorization to both Black+Good and White+Bad. As evidence of validity, parameter v was found to be related to explicit attitudes while the others were not (Klauer et al., 2007).

Parameter *a (response caution)*. This parameter describes the extent to which people prioritize accuracy or speed in decision-making (doing it right vs. doing it fast; Ratcliff et al., 2016). This speed-accuracy tradeoff is theorized to be a product of choices that people make ("I want to decide accurately" vs. "I want to decide quickly"). People using a conservative response mode respond more slowly but with high accuracy. People using a more liberal response mode respond more quickly at the risk of increased errors. People differ from each other in how much information they need before they make a decision (e.g., Voss et al., 2004). Parameter *a* quantifies this amount of evidence participants accumulate before making decisions. In addition, *a* is a function of task difficulty (Schmitz & Voss, 2012). In difficult tasks, people need more information before they feel confident enough to decide. The higher the *a*, the more information the participant sampled before deciding. Underpinning its ability to assess speed-accuracy tradeoffs, parameter *a* was found to be related to method-specific variance in the IAT (Klauer et al., 2007).

Parameter t_0 (non-decision-related processes). This parameter assesses all processes that precede and follow decision-making but are not involved in the actual decision-making process (Klauer et al., 2007). Thus, t_0 reflects the encoding of stimuli, response output processes, and motor responses (Lerche & Voss, 2017; Ratcliff et al., 2016; Voss et al., 2004). The higher the t_0, the more time the participant took to engage in these processes.

Applying Diffusion Models to the Understanding of Implicit Bias Reduction

Diffusion model analyses are well-suited to follow the request of Klauer et al. (2007) to investigate whether changes in IAT performances reflect changes in ease of decision-making, speed-accuracy tradeoffs or non-decision-related processes. Changes in overall IAT performance that have been observed in L2014 and L2016 could reflect changes in ease of decision-making (i.e., interventions changed participants ease in categorizing IAT stimuli), which could be attributed to changes in associations. However, they also could reflect changes in participants' response caution. They also could be a consequence of changes in

processes that do not relate to categorization-decisions at all but to components that precede and follow decisions (e.g., motor execution of key pressing in IATs). Diffusion models reveal insights into detailed processes that are related to changes in IAT performance.

The Present Study

We used diffusion models to investigate which processes were impacted by the interventions in the large-scale research projects of L2014 and L2016. Specifically, we investigated whether and how interventions that were designed to change *D*-scores affected ease of decision-making (IAT_v), speed-accuracy tradeoffs (IAT_a), and processes outside of decision-making (IAT_{t_0}).

Hypotheses

Immediate Intervention Effects

As changes in overall IAT performance reflect some form of psychological change, we predicted that the nine effective interventions in L2014 and L2016 would have different impacts on diffusion-model-based IAT effects. We predicted that the nine interventions which were ineffective at changing *D*-scores would be similarly ineffective at changing diffusion-model-based IAT effects.

Interventions Designed to Change Associations. IAT_v should increase due to changes in participants' information intake. Participants' ability to perform categorization-decisions and the ease of categorization should increase on incompatible trials and decrease on compatible trials (Table 1). IAT_a should decrease due to changes in the amount of information that is required by participants to come to a decision. The speed-accuracy tradeoff is a function of people's response mode and task difficulty. We predicted that participants would assign a comparable response mode to compatible and incompatible trials because due to interventions they should need a similar amount of information before feeling confident enough to decide. We predicted that IAT_{t_0} would *not* be impacted because the

interventions were developed to change the participants' actual decision processes (IAT_v and IAT_a) and not processes outside of it (IAT_{t_0}).

Faking the IAT. In line with recent studies (Röhner & Ewers, 2016b; Röhner & Thoss, 2018), we predicted faking to increase IAT_v, while decreasing IAT_a, and IAT_{t_0} (Table 1). Faking may involve establishing mental associations in order to fake that can impact the ease with which decisions are made (Ratcliff, 1978; Röhner & Ewers, 2016b). This results in an increased ability to perform categorization and increased decision-ease on incompatible trials while decreasing decision-ease on compatible ones (increased IAT_v). Faking on IATs involves deliberately adapting speed and accuracy (e.g., Fiedler & Bluemke, 2005). We predicted decreased IAT_a because participants may slow down responding and favor accuracy on compatible trials while they risk errors by trying to speed up on incompatible trials (Röhner et al., 2013). Parameter t_0 has been suggested to include the part of IAT performance that is related to behavior that goes against the instructions given for the measure (e.g., faking; Klauer et al., 2007). We predicted decreased IAT_{t_0} because participants may devote more time to processes that precede and follow actual decision tasks (e.g., taking more time to press response keys after compatible trials; Röhner et al., 2013).

Intervention Effects Over Time

L2016 found that *D*-scores returned to baseline after a delay. Therefore, we expected that the impacts of all interventions on IAT_v, IAT_a, and IAT_{t_0} would disappear as well.

Method

Data Sets

We reanalyzed data from two large-scale research projects (six studies) that examined (a) the impact of 18 interventions to change *D*-scores (L2014) and (b) the effectiveness of those interventions across time (L2016). 16,984 non-Black U.S. citizens/residents who volunteered for the study on the Project Implicit research website (http://implicit.harvard.edu) participated in and completed four studies in L2014: (Study 1:

N = 3,694; 66.1% women; 77.5% White; mean age 26.3; Study 2: *N* = 4,111; 65.3% women; 75.3% White; mean age 26.7; Study 3: *N* = 4,063; 67.7% women; 77.9% White; mean age 27.6; and Study 4: *N* = 5,116; 64.0% women; 77.2% White; mean age 31.3). 6,231 non-Black undergraduates participated in and completed two sessions in L2016: (Study 5[3]: *N* = 872 from Brock University and the University of Virginia; 72.9% women; 82.6% White; mean age 18.9; and

Study 6[4]: *N* = 4,888 from 17 American universities; 69.2% women; 60.8% White; mean age 19.2). The materials used in all studies are available for self-administration (http://osf.io/lw9e8/ and https://osf.io/um4ye/).

Measures

The seven-block IAT was used in Studies 1, 2, and 3 from L2014. Blocks 1, 2, and 5 (20, 20, and 40 trials) were practice blocks in which only two categories are presented. Participants are instructed to respond to exemplars of each category by pressing a key on the same side of the screen/keyboard as the category. In the first practice block, participants categorize images of Black faces and White faces. In the second practice block, they categorize Good and Bad words. In the third practice block, they again categorize Black faces and White faces, but the categories have switched sides. Blocks 3, 4, 6, and 7 form the critical blocks. Here, participants are asked to categorize stimuli from all four categories to two sides of the screen. In Blocks 3 and 4 (20 and 40 trials; compatible phase) of the race IAT, participants must respond to White images and Good words with one key and to Black images and Bad words with the other key. In Blocks 6 and 7 (20 and 40 trials; incompatible phase), participants must respond to Black images and Good words with one key and to White

[3] The study is called Study 1 in L2016. However, we changed it to Study 5 to avoid confusion because we already had another Study 1 from L2014.
[4] The study is called Study 2 in L2016. However, we changed it to Study 6 to avoid confusion because we already had another Study 2 from L2014.

images and Bad words with the other key. Both IAT phases were counterbalanced to control for order effects (Greenwald et al., 1998).[5]

The five-block IAT was used in L2014 Study 4 and in L2016 to reduce the length of each session. It was similar to the seven-block IAT except for two key changes. First, it had two critical blocks instead of four. Second, it had fewer trials (16 trials for the first two single blocks, 32 trials for critical blocks, and 24 trials for the single block between critical blocks).[6]

Interventions

The 18 interventions used to change *D*-scores in L2014 and L2016 were grouped into six categories (engaging with others' perspectives, exposure to counterstereotypical exemplars, appeals to egalitarian values, evaluative conditioning, inducing emotion, and intentional strategies to overcome biases). Between studies, some interventions were revised to improve their efficacy. Some interventions that did not work in earlier studies were excluded from later studies. We provide an overview of the interventions and their implementation in the studies hereafter. Our hypotheses were based on interventions that were effective at changing overall IAT performance.

Interventions that were effective in reducing *D*-scores

Exposure to counterstereotypical exemplars. Four out of five interventions that used experiences with counterstereotypical (positive Black and negative White) exemplars effectively reduced *D*-scores.

[1] Vivid counterstereotypical scenario. Based on research of Foroni and Mayr (2005) that showed changes in the performance on flower-insect IATs after they had presented fictional scenarios of dangerous flowers and good insects to participants, participants read a vivid second-person story in which they were the protagonist and were told to keep the story

[5] In addition, the position of the Good/Bad categories was randomized between-participants (half of the participants categorized Good to the left key and Bad to the right key, whereas the other half did the reverse).
[6] Like it was done for the seven-block IAT, block order was counterbalanced in the five-block IAT to avoid potential order effects.

in mind for the IAT. The story told participants to imagine walking down a street late at night after drinking at a bar. Suddenly, a White man in his 40s assaults the participant, throws him/her into the trunk of his car, and drives away. Some time passes before the White man opens the trunk and assaults the participant again. A young Black man notices the assault and saves the day by knocking out the White assailant.

In L2014, the length and vividness of the story increased from Study 1 to Study 2 (e.g., from "With sadistic pleasure, he bashes you with his bat again and again" to "With sadistic pleasure, he beats you again and again. First to the body, then to the head. You fight to keep your eyes open and your hands up. The last things you remember are the faint smells of alcohol and chewing tobacco and his wicked grin"). In L2014 Study 3, the instructions to affirm positive Black associations and negative White associations were revised to include two sets of pictures. One set showed the stimuli for Black people on the IAT paired with the word Good, whereas the other set showed the stimuli for White people on the IAT paired with the word Bad. In L2014 Study 4 and the L2016 studies, only one set of pictures was included.

[2] Practicing an IAT with counterstereotypical exemplars. Prior research had shown that the exposure to positive Blacks and negative White persons changes IAT performance (e.g., Joy-Gaba & Nosek, 2010). In this intervention, participants saw pictures of famous positive Black (e.g., Oprah Winfrey) and infamous negative White (e.g., Adolf Hitler) exemplars along with brief one-line descriptions of what they are known for. After this, participants were asked to complete part of an IAT with combined blocks. These blocks consisted of the same stimuli that were used in the race IAT, along with six positive Black and six negative White exemplars.

Due to a programming error in L2014 Study 1, participants learned they were going to take part in a race IAT and saw the positive Black and negative White exemplars that would accompany the standard Black and White images. However, they did not complete the counterstereotypical practice. Due to this error, data from Study 1 were not analyzed. For

L2014 Study 2, the procedure was implemented as described above, with the combined block consisting of 90 trials. L2014 Studies 3 and 4 reduced the number of trials in the combined block to 52 trials, and L2016 further reduced it to 32 trials. In all studies in L2014 and in L2016 Study 5, the same stimuli were applied. In L2016 Study 6 some of the negative White exemplars were replaced with more recent exemplars (e.g., Bernie Madoff) because the older exemplars may not have been familiar to undergraduates in 2014 (e.g., John Gotti).

[3] Shifting group boundaries through competition. Research showed that intense competition and strong outgroup threats lead to negative outgroup attitudes (Riek et al., 2006). The idea was that cooperating with racial outgroup members to compete the ingroup members could change *D*-scores. Thus, participants played a simulated dodgeball game in which their teammates were Black and their opponents were White. Whereas the Black teammates saved the participants from being knocked out and were good sports, the White opponents played unfairly and were bad sports. Subsequent, participants were asked to think "Black = Good" and "White = Bad" and to remember how their Black teammates helped them and their White opponents hurt them while completing the IAT. This intervention was first applied in L2014 Study 2. In order to adhere to time constraints, sections requiring participants' input were set to automatically advance if participants responded too slowly in L2014 Studies 3 and 4 and in both L2016 studies.

[4] Shifting group affiliations under threat.

Outgroup threats lead to more negative outgroup attitudes (Riek et al., 2006). Flipping the group memberships may have the opposite effect. After participants read a vivid and threatening post-nuclear war scenario, they were shown profiles of people described as "close friends" in their camp and "terrible enemies." All of the "close friends" were Black and had helpful survival skills (e.g., doctor), whereas all of the "terrible enemies" were villainous White people who wanted to destroy the camp. After having read the profiles, participants

were asked: “Please imagine and think about the friends and enemies you just read about while you complete these tasks.”

This intervention was first applied in L2014 Study 2 and included only “close friends” profiles. L2014 Study 3 added “terrible enemies” profiles. L2014 Study 4 changed the faces of Black individuals to be more likable and the faces of White individuals to be less likable. This intervention was also used in both L2016 studies.

Appeals to egalitarian values. One out of five interventions that used appeals to egalitarian values effectively reduced *D*-scores.

[5] Priming multiculturalism. Priming multiculturalism, the ideology that racial differences should be acknowledged and celebrated, had previously been shown to reduce race IAT *D*-scores (Richeson & Nussbaum, 2004). This intervention adapted Richeson and Nussbaum’s (2004) approach. After reading a prompt that advocated multiculturalism, participants were asked to summarize the content in their own words and to list two reasons why multiculturalism “is a positive approach to interethnic relations.” They were also instructed to think “Black = Good” on the subsequent IAT. This intervention was first applied in L2014 Studies 3 and 4. It was kept in both L2016 studies.

Evaluative conditioning. Both interventions that used evaluative conditioning effectively reduced *D*-scores.

[6] Evaluative conditioning. The idea behind evaluative conditioning is to shift the attitude of a first object toward the direction of a second, valenced object by presenting both objects together (Olson & Fazio, 2006). Thus, participants were presented Black people’s faces paired with positive words and White people’s faces paired with negative words. Each picture-word pair was presented one at a time in the center of their computer screen for 1 s. The stimuli were the same faces and words that were used in the subsequent IAT. Participants’ task was to categorize the face as either Black or White by pressing the E key or I key. Participants were also instructed to memorize the presented words for a subsequent test

at the end of the categorization task. In L2014 Study 1, this intervention consisted of 48 trials of paired stimuli. It was reduced to 40 trials in L2014 Study 3, L2014 Study 4, and in L2016. In L2014 Study 2, participants did not complete the recall task.

[7] Evaluative conditioning with the GNAT. Participants completed a version of the Go/No-Go Association Task (GNAT; Nosek & Banaji, 2001). The idea behind this approach was to strengthen associations between Black and Good and White and Bad. The stimuli in this task included pictures of Black and White people and Good and Bad words. Participants were asked to respond to picture-word pairings that were presented on the computer screen one at a time by pressing the space bar when the stimulus pair matched two categories and by refraining from pressing it when there was no match. In the first GNAT block, the majority of stimulus pairings consisted of pictures of Black people and Good words. Participants were asked to press the space bar when the stimulus pair consisted of a picture of a Black person and a Good word and not to press it for all other stimulus pairings. In the second GNAT block, a minority of stimulus pairings consisted of pictures of White people and Good words. Participants were asked to press the space bar when the stimulus pair consisted of a picture of a White person and a Good word and not to press it for all other stimulus pairings. In L2014 Study 1, this intervention consisted of 100 trials of paired stimuli. In L2014 Study 2, the number of trials was reduced to 60. Also, the "go" category for both blocks was "Black and Good," and the second block of trials required a faster response than the first did. In L2014 Study 3 and both L2016 studies most of these features were retained but the number of trials was reduced to 45. Participants were also instructed to count the number of times pictures of Black people were paired with Good words over the course of the task.

Intentional strategies to overcome biases. Both interventions that gave participants strategies to control their associations effectively reduced *D*-scores.

[8] Using implementation intentions. Implementation intentions can be used to make associations between behavior and cues more accessible in memory and therefore increase

consistency between intentions and behavior (Brandstätter et al., 2001). Participants completed a short tutorial on how to take an IAT. This tutorial informed them that people who complete the IAT tend to exhibit associations preferring White people relative to Black people. Afterwards, participants were instructed to commit themselves to an implementation intention (an "if-then" plan that ties a behavioral response to a situational cue; Gollwitzer, 1999) by saying to themselves silently, "I definitely want to respond to the Black face by thinking 'good.'"

In L2014 Study 1, this intervention proceeded as described above. In all other studies of L2014 and L2016 participants completed practice IAT trials before being given the implementation intention instructions.

[9] Faking the IAT. Faking changes *D*-scores (e.g., Röhner et al., 2011). Participants completed a tutorial on how to take the IAT. In this tutorial, they were informed that people who complete the IAT tend to exhibit associations preferring White people relative to Black people. Afterwards, participants were told that they were participating in a study on faking the IAT. They were given empirically supported faking strategies (Röhner et al., 2013) that asked participants to slow down on blocks in which "Black and Bad" were paired and accelerate on blocks in which "White and Bad" were paired. Participants were instructed to ignore instructions on the IAT that contradicted faking instructions. In L2014 Study 1, this intervention proceeded as described above. In all other studies of L2014 and L2016, participants additionally completed IAT practice trials before being instructed in how to fake.

Interventions that were ineffective at reducing *D*-scores

Engaging with others' perspectives. None of the three interventions that used engaging with others' perspectives effectively reduced *D*-scores.

[10] Training empathic responding. Prior research by Finlay and Stephan (2000) demonstrated that an intervention to increase empathy toward Black people reduced racial prejudice. This intervention tried to increase empathy toward Black people with a game in

which participants were shown images of Black people expressing happiness, sadness, angriness, or fear. After seeing the image participants were asked to identify the emotion that best described the emotion they saw. They were also asked to select the reason the portrayed person was feeling this way (e.g., for fear: "A snake crawled up my leg"). If participants selected the correct emotion and rationale, they were rewarded with a smiling face and the phrase "Thanks for understanding." In L2014 Study 1, participants chose from four response options for both emotion identification and emotion reason questions. In L2014 Study 2, they chose from two options. In L2014 Studies 3 and 4 and in L2016, this intervention was not tested.

[11] Perspective taking. Perspective-taking can change *D*-scores through positive associations between the self and an outgroup member (Todd & Burgmer, 2013). Participants in this intervention were presented five scenarios with pictures of Black people. Each scenario was accompanied by an emotional context (e.g., "This person is getting married in the morning"). Participants' task was it to imagine that they were the portrayed person and write down how they felt. This intervention was tested only in L2014 Study 1.

[12] Imagining interracial contact. Prior research found that participants who imagined contact with outgroup members showed more positive attitudes toward them afterward (e.g., Turner & Crisp, 2010). Participants were asked to imagine interacting with a Black stranger in a relaxed, positive, and comfortable environment and to list as many details as possible about the imagined interaction (L2014, Study 1). In L2014 Study 2, participants were additionally asked to imagine a negative interaction with a White person and to list as many details as possible. The corresponding prompts came along with a photograph of a smiling Black woman and a photograph of a frowning White woman. This intervention was not tested in L2014 Studies 3 and 4 or in L2016.

Exposure to counterstereotypical exemplars. One of the five interventions that gave participants experience with counterstereotypical exemplars did not effectively reduce *D*-scores.

[13] Highlighting the value of a subgroup in competition. This intervention was based on the ingroup identity model that predicts an emphasis on superordinate identities will biases toward outgroup members (Gaertner & Dovidio, 2000). Participants read a description of an international competition in basketball. The description declared that the United States has one of the most successful basketball teams in the world and is expecting intense competition from other countries. A list with the names of eight prominent basketball players that were predominantly Black was presented to participants, who were asked to mark all the names they recognized. The questionnaire was designed to indirectly remind participants that most American basketball players are Black. This intervention was only tested in L2014 Study 1.

Appeals to egalitarian values. Four out of five interventions that used appeals to egalitarian values did not effectively reduce *D*-scores.

[14] Priming feelings of nonobjectivity. This intervention aimed to make participants aware that they could behave in a biased manner, so that they would be more motivated to override their biases on the IAT. In L2014 Study 1, participants were asked to remember nine examples from their past in which they behaved objectively. In L2014 Study 2, participants stated *how they would personally act* in a particular decision and *how they think society believes they should act* when making this decision. In L2014 Studies 3 and 4, participants were asked to read a popular science article dealing with psychological biases outside of conscious awareness and its potential impact on behavior. This intervention was not tested in L2016.

[15] Considering racial injustice. Explicit racial attitudes toward outgroups can be improved by reflecting on egalitarian values (Katz & Hass, 1988). This intervention examined whether reflecting on egalitarian values could affect implicit racial attitudes.

Participants were asked to write down examples of injustices that White people have perpetrated on Black people in the past, examples of injustices that White people currently perpetrate on Black people, and examples of ways in which Black people have overcome racial injustice. In L2014 Study 1, participants listed two of each of these examples. In L2014 Study 2, it was reduced to one of each example. This intervention was not tested in L2014 Studies 3 or 4 or in L2016.

[16] Instilling a sense of common humanity. Expanding the boundaries of one's ingroup can reduce explicit intergroup biases (Gaertner et al., 1993). This intervention used a video-based intervention to see if the same could happen with implicit racial attitudes. Participants were asked to watch a video that shows a man dancing with people in different countries all over the world (http://www.youtube.com/watch?v=zlfKdbWwruY). This intervention was first applied in L2014 Study 2 and was kept through Study 4. It was not tested in L2016.

[17] Priming an egalitarian mindset. This intervention tried to prime an egalitarian ideology and was based on research that demonstrated that priming social ideologies can shift racial attitudes (Sears & Henry, 2005). To prime an egalitarian mindset, participants filled out the Humanitarian-Egalitarianism scale (Katz & Hass, 1988) in L2014 Study 1. In Study 2, they wrote a short essay in support of the declaration, "All people and groups are equal; therefore, they should be treated the same way." In Studies 3 and 4, participants wrote about a time they failed to live up to egalitarian ideals after filling out a questionnaire that asked them how important it was to be egalitarian. This intervention was not tested in L2016.

Inducing emotion. The intervention that induced moral elevation did not effectively reduce *D*-scores.

[18] Inducing moral elevation. Witnessing acts of charity, gratitude, or generosity can induce the emotion of "elevation" (e.g., Haidt, 2003) that may reduce prejudice by blurring boundaries between the ingroup and the outgroup. In order to induce moral elevation,

participants watched a video about a high school girls' softball game in L2014 Study 1. The video showed White players exhibiting extraordinary sportsmanship by carrying an opposing White player around the bases after she injured herself as she hit a homerun. In L2014 Study 2, participants watched a video showing a Black high school music teacher who expressed his gratitude toward his former music teacher (also Black) because he had seen promise in him when he was a teenager and thus, saved him from a life of crime. This intervention was not tested in L2014 Studies 3 or 4 or in L2016.

Analytical Approach

To conduct diffusion models, we used the Excel-based EZ software from http://www.ejwagenmakers.com/papers.html. The EZ diffusion model has already been applied successfully to IAT data (e.g., Röhner & Thoss, 2018) and has demonstrated high statistical power to detect effects even in small samples (e.g., Van Ravenzwaaij et al., 2016). We followed conventions and set the arbitrary scaling parameter *s* to 0.1.

Pretreatment of data sets. We followed the recommendation made by Voss et al. (2013) as well as Voss and Voss (2008) to remove outliers from the individual response-time distribution if participants had reaction times below 200 ms or above 5,000 ms. Altogether we excluded 15,023 trials (0.5% of the trials).[7]

Computation of input variables. We computed mean reaction times, variance of reaction times, and percentage of correct responses separately for each participant (*N*s = 3,591, 4,009, and 1,999 in Studies 1, 2, and 3; *N*s = 5,054, 1,021, and 5,295 for posttest IATs in Studies 4, 5, and 6; *N*s = 1,257, and 6,022 for follow-up IATs in Studies 5, and 6) and each combined IAT phase type (compatible vs. incompatible) within every measurement occasion (depending on the study, up to two measurements: posttest and follow-up).

[7] We excluded 1,555 trials from the IAT (0.4% of the trials) in Study 1, 1,667 trials (0.3% of the trials) in Study 2, 759 trials (0.3% of the trials) in Study 3, and 1,312 trials (0.4% of the trials) from the posttest IAT in Study 4 of the L2014 studies. We excluded 112 trials (0.2% of the trials) from the posttest IAT, and 963 trials (1.2% of the trials) from the follow-up IAT in Study 5, and we excluded 614 trials (0.2% of the trials) from the posttest IAT, and 5,582 trials (1.5% of the trials) from the follow-up IAT in Study 6 of the L2016 studies.

Following Wagenmakers et al.'s (2007) recommendation, we corrected participants' percentage of correct responses that equaled exactly 1.0 by subtracting half an error from the percentage of correct responses before running further analyses. We also corrected participants' percentage of correct responses that equaled exactly 0 and 0.5 by adding half an error, respectively. Because of the approximation formula, $t0$ can become negative in sign (e.g., the mean of the reaction time is less than the mean decision time that is defined: $\frac{a}{2v} \times \frac{1-e^{y}}{1 \mp e^{y}}$; Wagenmakers et al., 2007). However, a negative $t0$ cannot be interpreted theoretically because it represents the non-decisional portion of response time and time cannot take on negative values (Voss et al., 2004). Thus, it is recommended to exclude participants with negative $t0$ before further analyses (Wagenmakers et al., 2007). Altogether, we excluded N = 1,416 (5% of participants) from further analyses because $t0$ became negative in sign[8]. Finally we excluded participants that had more trials than the typical amount due to a technical issue during data collection.[9] The final number of participants across the several conditions that were included in our analyses are shown in the Supplements at the OSF (https://osf.io/9xegp/?view_only=0baa541b0c164899b7dbed8f4e32e59a).

Parameter estimation. We used EZ to estimate independent diffusion models for each participant and each combined IAT phase type within every measurement occasion. Altogether, we computed 53,648 EZ diffusion models (Study 1: 3,378 participants; Study 2: 3,767 participants; Study 3: 1,902 participants; Study 4: 4,776 participants; Study 5 posttest

[8] (N = 209 [5.8%] participants from Study 1, N = 237 [6.6%] from Study 2, N = 96 [2.7%] from Study 3, and N = 275 [7.7%] from Study 4 regarding the posttest IAT; N = 33 [0.9%] from the posttest IAT, and N = 39 [1.1%] from the follow-up IAT in Study 5, and, N = 197 [5.5%] from the posttest IAT, and N = 205 [5.7%] from the follow-up IAT in L2016 Study 6) or because the IDs had less trials than the typical amount (participants did not finish the IAT; N = 0 [0.0%] participants from Study 1, N = 0 [0.0%] from Study 2, N = 0 [0.0%] from Study 3, and N = 0 [0.0%] from Study 4 regarding the posttest IAT; N = 5 [0.1%] from the posttest IAT, and N = 22 [0.6%] from the follow-up IAT in Study 5, and, N = 13 [0.4%] from the posttest IAT, and N = 53 [1.5%] from the follow-up IAT in L2016 Study 6)

[9] (N = 4 [0.1%] participants from Study 1, N = 5 [0.1%] from Study 2, N = 1 [0.0%] from Study 3, and N = 3 [0.1%] from Study 4 regarding the posttest IAT; N = 1 [0.0%] from the posttest IAT, and N = 1 [0.0%] from the follow-up IAT in Study 5, and, N = 2 [0.1%] from the posttest IAT, and N = 15 [0.4%] from the follow-up IAT in L2016 Study 6)

IAT: 982 participants; Study 5 follow-up IAT: 1,195 participants; Study 6 posttest IAT: 5,083 participants; Study 6 follow-up IAT: 5,741 participants). For the seven-block IAT, each diffusion model analysis was based on an average of 118.6 trials after outliers were excluded. For the five-block IAT, each diffusion model analysis was based on an average of 63.4 trials after outlier exclusions.

Computation of diffusion-model-based IAT effects. Using the parameters estimated in EZ diffusion model analyses, we computed IAT_v, IAT_a, and IAT_{t_0} (Röhner & Thoss, 2018). We did so by subtracting the parameters that were estimated on compatible trials from the parameters that were estimated on incompatible trials (Table 1).

Analyzing diffusion-model-based IAT effects. We used paired-sample *t* tests and meta-analyses to test our hypotheses. For all analyses, we set α to .05. To be concise, we focus on presenting the meta-analytic results whenever possible. Descriptive analyses and significance tests by study, condition, and IAT effect are available in the Supplements at the OSF (https://osf.io/9xegp/?view_only=0baa541b0c164899b7dbed8f4e32e59a).

Results

Control Condition

In the control conditions decision-making (IAT_v) was easier for participants on compatible trials and more difficult on incompatible trials, *d*s = -0.99, -0.92, -1.07, and -0.84, in Studies 1, 2, 3, and 4; *d*s = -0.93, -0.92, -0.96, and -1.06; for the posttest IAT, and follow-up IAT in Studies 5 and 6, respectively. In addition, participants in the control condition responded more cautiously on the "more difficult" incompatible trials and responded less cautiously on the "easier" compatible trials (IAT_a), *d*s = 0.34, 0.35, 0.40, and 0.28, in Studies 1, 2, 3, and 4; *d*s = 0.49, 0.35, 0.52, and 0.58; for the posttest IAT, and follow-up IAT in Studies 5 and 6, respectively. Lastly, participants in the control condition spent more time on processes outside actual decision processes on incompatible trials than on compatible trials

(IAT_{t_0}), *d*s = 0.17, 0.10, 0.19, and 0.36, in Studies 1, 2, 3, and 4; *d*s = 0.40, 0.28, 0.14, and 0.13; for the, posttest IAT, and follow-up IAT in Studies 5 and 6, respectively.

Impacts of Interventions on Diffusion-Model-Based IAT Effects

We compared the impact of interventions on *D*-scores to the impact of interventions on the three diffusion-model-based IAT effects: ease of decision-making (IAT_v), response caution (IAT_a), and non-decision-related processes (IAT_{t_0}). We present the results of the meta-analyses within the text and in Figure 1. Mean IAT effects for each condition across the six studies from L2014 and L2016 as well as the results at a study-by-study-level are shown in the Supplements at the OSF (https://osf.io/9xegp/?view_only=0baa541b0c164899b7dbed8f4e32e59a).

Relation Between Change on *D*-Scores and Change on Diffusion-Model-Based IAT Effects

To assess correspondence between *D*-scores and diffusion-model-based IAT effects, we calculated Spearman's rank correlation coefficients between the ranked (descriptive) effectiveness of each intervention on the *D*-score and each of the diffusion-model-based IAT effects (1=most effective; 18=least effective). Interventions that had the strongest effects on reducing *D*-scores, also had the strongest effects on increasing IAT_v, *r*s(16) = .90, $p \leq .001$, and reducing IAT_a, *r*s(16) = .85, $p \leq .001$. The extent to which an intervention reduced *D*-scores was unrelated to its influence on IAT_{t_0}, *r*s(16) = .25, $p = .324$.

Immediate Intervention Effects

Interventions that were effective in reducing *D*-scores

Exposure to counterstereotypical exemplars. The four interventions that exposed participants to counterstereotypical exemplars and reduced *D*-scores (vivid counterstereotypical scenario, practicing an IAT with counterstereotypical exemplars, shifting group boundaries through competition, and shifting group boundaries under threat) all increased IAT_v, *d*s and 95% CIs = 0.38 [0.32, 0.45], 0.21 [0.14, 0.27], 0.30 [0.23, 0.37], 0.22

CI [0.14, 0.29], decreased IAT_a, *d*s and 95% CIs = -0.32 [-0.395, -0.25], -0.30 [-0.37, -0.23], -0.25 [-0.33, -0.18], -0.14 95% CI [-0.22, -0.07], and did not affect IAT_{t_0}, *d*s and 95% CIs = 0.01 [-0.06, 0.01], 0.02 [-0.05, 0.09], 0.01 [-0.09, 0.06], 0.00 [0.07, 0.08].

Appeals to egalitarian values. The intervention that appealed to egalitarian values and changed *D*-scores (priming multiculturalism) increased IAT_v, $d = 0.18$, 95% CI [0.09, 0.26], and decreased IAT_a, $d = -0.14$, 95% CI [-0.22, -0.06], but did not impact IAT_{t_0}, $d = -0.05$, 95% CI [-0.13, 0.03].

Evaluative conditioning. The two interventions that used evaluative conditioning and changed *D*-scores (evaluative conditioning and evaluative conditioning with the GNAT) increased IAT_v, *d*s and 95% CIs = 0.18 [0.11, 0.25], 0.20 [0.13, 0.27] and decreased IAT_a, *d*s and 95% CIs = $d = -0.14$ [-0.21, -0.07], -0.15 [-0.22, -0.08]. Evaluative conditioning with the GNAT slightly decreased IAT_{t_0}, $d = -0.08$ 95% CIs [-0.51, -0.01]. Evaluative conditioning did not affect IAT_{t_0}, $d = 0.00$ 95% CIs [-0.07, 0.07].

Intentional strategies to overcome biases. The two interventions that used intentional strategies to overcome biases and changed *D*-scores (implementation intentions and faking) increased IAT_v, *d*s and 95% CIs = 0.33 [0.26, 0.39], 0.30 [0.23, 0.37], and decreased IAT_a, *d*s and 95% CIs = -0.27 [-0.34, -0.20], -0.52 [-0.59, -0.45]. Faking decreased IAT_{t_0}, $d = -0.13$ 95% CI [-0.20, -0.07], but using implementation intentions did not, $d = 0.00$ 95% CI [-0.07, 0.07].

Interventions that were not effective in reducing *D*-scores

Engaging with others' perspectives.

None of the interventions that included engaging with others' perspectives (empathy training, perspective-taking, and imagining interracial contact) affected IAT_v, *d*s and 95% CIs = 0.08 [0.04, 0.20], 0.00 [-0.18, 0.18], 0.00 [-0.12, 0.12], IAT_a, *d*s and 95% CIs = 0.00 [-0.12,

0.12], 0.00 [-0.18, 0.18], -0.12 [-0.24, 0.01], or IAT_{t_0}, *d*s and 95% CIs = 0.06 [-0.06, 0.19], -0.07 CI [-0.25, 0.11], 0.06 [-0.06, 0.19].

Exposure to counterstereotypical exemplars. The one intervention that exposed participants to counterstereotypical exemplars and did not affect the *D*-score (highlighting the value of a subgroup in competition) did not affect IAT_v, d = 0.00, 95% CI [-0.17, 0.17], IAT_a, d = -0.25, 95% CI [-0.42, 0.08], or IAT_{t_0}, d = 0.12, 95% CI [-0.04, 0.29], either.

Appeals to egalitarian values. With the exception of considering racial injustice that increased IAT_v, *d* and 95% CI = 0.15 [0.03, 0.28], interventions that appealed to egalitarian values and were ineffective at reducing *D*-scores (priming feelings of nonobjectivity, instilling a sense of common humanity, and priming an egalitarian mindset) did not affect IAT_v, *d*s and 95% CIs = -0.05 [-0.13, 0.04], -0.04 [-0.14, 0.06], 0.03 [-0.16, 0.11]. All four interventions did not affect IAT_a, *d*s and 95% CIs = 0.00 [-0.08, 0.08], 0.00 [-0.12, 0.12], 0.11 [-0.01, 0.21], 0.07 [-0.01, 0.16] or IAT_{t_0}, *d*s and 95% CIs = 0.00 [-0.08, 0.09], 0.03 [-0.09, 0.15], -0.08 [-0.81, 0.01], 0.03 [-0.05, 0.12].

Inducing Emotion. Inducing emotions in terms of moral elevation did not affect IAT_v, d = 0.07, 95% CI [-0.06, 0.20], IAT_a, d = -0.23, 95% CI [-0.36, 0.10], or IAT_{t_0}, d = 0.13, 95% CI [0.00, 0.26].

Intervention Effects Over Time

The impacts of interventions on *D*-scores declined over time[10] in L2014 and L2016. Similarly, the impacts of all of the interventions on the ease with which participants make their decisions (IAT_v) all disappeared over time as well (*d*s and 95% CIs = vivid

[10] With the exception of a small but significant effect of implementation intentions on *D*-scores, d = -0.12, 95% CI [-0.24, -0.08], there were no significant effects of any of the interventions across time (ds and 95% CIs = vivid counterstereotypical: -0.06 [-0.18, 0.06]; practicing an IAT with counterstereotypical exemplars: -0.03 [-0.15, 0.08]; shifting group boundaries through competition: -0.09 [-0.21, 0.02], shifting group boundaries under threat: -0.02 [-0.14, 0.09]; priming multiculturalism: -0.05 [-0.17, 0.06]; evaluative conditioning: -0.01 [-0.12, 0.11]; evaluative conditioning with the GNAT: -0.09 [-0.21, 0.03], or faking: -0.03 [-0.15, 0.08]). The small but significant effect of implementation intentions on *D*-scores can be explained as follows: In the original L2016 paper, the authors set a critical *p*-value of .01 and analyzed the two studies separately. When the two studies were combined meta-analytically for this article, the effect was statistically significant at $p < .01$.

counterstereotypical scenario: -0.01 [-0.13, 0.11]; practicing an IAT with counterstereotypical exemplars: 0.04 [-0.08, 0.16]; shifting group boundaries through competition: 0.05 [-0.06, 0.17], shifting group boundaries under threat: 0.00 [-0.12, 0.12]; priming multiculturalism: 0.05 [-0.07, 0.17]; evaluative conditioning: -0.03 [-0.15, 0.09]; evaluative conditioning with the GNAT: 0.07 [-0.05, 0.19]; using implementation intentions: 0.09 [-0.03, 0.21]; faking: -0.02 [-0.14, 0.10]).

There was a small but significant impact of implementation intentions on IAT_a, d = -0.14 95% CI [-0.25, -0.02]. The impacts on IAT_a disappeared for all other interventions (ds and 95% CIs = vivid counterstereotypical scenario: 0.01 [-0.11, 0.13]; practicing an IAT with counterstereotypical exemplars: -0.11 [-0.22, 0.01]; shifting group boundaries through competition: -0.09 [-0.20, 0.03], shifting group boundaries under threat: 0.00, [-0.12, 0.12]; priming multiculturalism: -0.07 [-0.19, 0.05]; evaluative conditioning: -0.07 [-0.19, 0.05]; evaluative conditioning with the GNAT: -0.12 [-0.24, 0.00]; faking: -0.02 [-0.13, 0.10]).

None of the interventions had an impact on non-decision-related processes (IAT_{t_0}) across time (ds and 95% CIs = vivid counterstereotypical scenario: -0.06 [-0.18, 0.05]; practicing an IAT with counterstereotypical exemplars: 0.01 [-0.11, 0.12]; shifting group boundaries through competition: 0.01 [-0.11, 0.13]; shifting group boundaries under threat: -0.01 [-0.13, 0.11]; priming multiculturalism: -0.03 [-0.15, 0.09]; evaluative conditioning: -0.01 [-0.13, 0.11]; evaluative conditioning with the GNAT: -0.03 [-0.16, 0.09]; using implementation intentions: 0.03 [-0.08, 0.15]; faking: -0.09 [-0.21, 0.03]).

General Discussion

Using diffusion modelling, we examined the mental processes changed by 18 interventions tested in L2014 and L2016. We found that the nine interventions that reduced D-scores changed ease of decision-making (IAT_v) such that deciding on incompatible trials became easier and deciding on compatible trials became more difficult. This is suggestive of a reduction in White+Good/Black+Bad associations. Effective interventions also impacted

speed-accuracy tradeoffs (IAT_a) such that participants made more similar tradeoffs between speed vs. accuracy on compatible and incompatible trials. Only faking on the IAT and evaluative conditioning with the GNAT affected non-decision-related processes (IAT_{t_0}). Finally, we found little evidence for intervention effects on IAT_v, IAT_a, or IAT_{t_0} over longer periods of time.

Interventions Temporarily Changed Ease of Decision-Making

Our results go beyond L2014 and L2016 by offering insight into the processes *behind* implicit bias reductions. IAT_v reflects participant's ability to perform the categorization and the ease of the task (Klauer, 2014). Interventions that effectively changed overall IAT performance changed how easily participants responded to trials (IAT_v). They most likely reduced the activation of White+Good/Black+Bad associations, which in turn impacted the ease of decision-making. Thus, participants in effective intervention condition collected information on incompatible trials more quickly and with fewer errors (with more ease), but were slower and made more errors on compatible trials than participants in control conditions.

How might changes in the ease of decision-making reflect changes in associations? Effective interventions may have motivated and/or enabled participants to invest more processing capacity and/or attention to incompatible trials than to compatible trials (see Klauer, 2014), which then facilitated decision processes on incompatible trials (increasing IAT_v). Another possibility is related to the ease of the task that is reflected in IAT_v. Effective interventions may have changed how people interpreted the stimuli (Klauer, 2014) in a way that compatible trials became less easy and incompatible trials became easier for them.

Interventions Temporarily Changed Speed-Accuracy Tradeoffs

Interventions that effectively reduced *D*-scores also decreased IAT_a. Increases in task-difficulty lead to a more cautious response style (Schmitz & Voss, 2012). In line with this finding, participants in control conditions showed a more conservative response mode on more difficult (incompatible) trials than on easier (compatible) trials. However, participants

that completed the interventions that were effective at reducing *D*-scores applied a nearly equal speed-accuracy tradeoff on compatible and incompatible trials. Interventions might have motivated people to be more consistent across trials in terms of raising concerns of racial fairness, which might account for the observation of devoting a comparable amount of response caution on both type of IAT trials.

The robust findings of changes on IAT_v and on IAT_a may be related. As the increased ease of decision-making on incompatible trials and increased difficulty on compatible trials may result from participants' increased ability and/or motivation to invest cognitive resources on incompatible trials, a reduction in differences in speed-accuracy tradeoffs may be because participants perceive task difficulty more similarly between compatible and incompatible trials.

Two Interventions Temporarily Affected Non-Decision-Related Processes

Only faking on the IAT (d = -0.13) and evaluative conditioning with the GNAT (d = -0.08) decreased IAT_{t_0}. Supporting earlier theory and evidence (Klauer et al., 2007; Röhner et al., 2013; Röhner & Thoss, 2018), participants in the faking condition took more time before and after categorization on compatible trials relative to incompatible trials compared to the control condition. Unexpectedly, participants who took the evaluative conditioning with the GNAT showed similar behavior (albeit to a lesser extent). This may be due to interference from task-switching. In the intervention, participants pressed a button when a particular pairing appeared and abstained from pressing it when the different kind of pairing appeared. In the IAT, participants sorted stimuli into categories. Mental interference from the intervention's prior task instructions may have impeded processes outside of decision-making on the IAT.

No Intervention Effects over Time

Like the effects on *D*-scores (L2016), the effects on ease of decision-making were only temporary, rebounding within several days. Thus, the psychological changes that

increased ease of categorization either disappeared or were overwritten (e.g., by information within daily routine). Effects of interventions on speed-accuracy tradeoffs disappeared over time with the exception of a small effect of implementation intentions (d = -0.14). The finding of implementation effects having a somewhat more durable impact on speed-accuracy tradeoffs mirrors the small effect of implementation intentions on D-scores (d = -0.12) that was significant at the $\alpha = .05$ but not $\alpha = .01$ level (L2016, Study 2). Although this effect persisted, the size of the effect was considerably smaller than the effects observed immediately after intervention. The effects of all interventions on non-decision-related processes disappeared over time. Like the effects on IAT_a, the effects on IAT_{t_0} disappeared as the impact on IAT_v disappeared.

Insights into Intervention Effectiveness

Interventions strongly varied in their respective approach. In this section, we focus on the relative size of intervention effects and what those effects mean for interpretation. In terms of changing ease of decision-making, the effective interventions ranked from most to least effective by meta-analytic effect size were: *vivid counterstereotypic scenario, using implementation intentions, shifting group boundaries through competition, faking, shifting group affiliations under threat, practicing an IAT with counterstereotypical exemplars, evaluative conditioning with the GNAT, evaluative conditioning, priming multiculturalism*, and *considering racial injustice*. The rest of the interventions did not significantly affect ease of decision-making. Interventions that gave participants experiences with counterstereotypical people or intentional strategies to override bias affected participants ease in decision making more than interventions that used evaluative conditioning or asked people to reflect on egalitarian values. The most effective interventions tended to be more vivid (e.g., shifting group affiliations under threat), emotional (e.g., vivid counterstereotypic scenario), active (e.g., implementation intentions), or self-relevant (e.g., shifting group boundaries under competition). However, we note that range in effect size among effective

interventions (*d*'s = 0.15 to 0.38) was smaller than range of *D*-score effect sizes in L2014 an L2016 (*d*'s = .18 to .58). This may be because IAT_v reflects only one part of overall IAT performance (*D*-score).

In terms of changing response caution, the effective interventions ranked from most to least effective by meta-analytic effect size were: *faking, vivid counterstereotypic scenario, practicing an IAT with counterstereotypical exemplars, using implementation intentions, shifting group boundaries through competition*, *highlighting the value of a subgroup in competition, inducing moral elevation*, *evaluative conditioning with the GNAT, shifting group affiliations under threat, priming multiculturalism,* and *evaluative conditioning*. The rest of the interventions did not significantly affect response caution. Among effective interventions faking (*d*'s = -0.52) popped out having a medium effect size while the other effective interventions had small effect sizes (*d*'s = -0.32 to -0.14). Faking may have been more effective than the rest at changing response caution because the faking manipulation involved the deliberate and intentional choice to adapt speed and accuracy (Röhner & Ewers, 2016). For the rest of the interventions, changes in response caution may have been a side effect of changes in ease in decision-making.

Only *faking,* and *evaluative conditioning with the GNAT* effectively changed non-decision related processes (*d*'s = -0.13 and -0.08). Thus, interventions rarely affected processes outside of the actual decision process. While the small impact on non-decision-related processes of evaluative conditioning with the GNAT may result from mental interference from the intervention's prior task, faking was expected to impact behavior outside of the decision process because of the deliberate use of strategies that goes against the task instructions (Klauer et al., 2007; Röhner & Thoss, 2018).

Note on Single-, Dual-, and Many-Process Model Perspectives of Evaluation

The results shed light on the issue of which processes are changed by interventions that impact overall IAT performance. One might ask whether the fact that the results of

L2014 and L2016 can be explained with diffusion models challenges the traditional view of dual processes in social cognition (e.g., Strack & Deutsch, 2004). In diffusion models, associative and reflective influences are mirrored in one unitary decision process. However, it is important to note that being able to analyze and interpret data with a model that suggests one decision-making process does not necessarily mean that the cognitive structure underlying these decisions represents a single process. Just as with *D*-scores, our results on diffusion-model-based IAT effects are compatible with single-, dual- and many-process perspectives of evaluation and cannot cleanly arbitrate between these perspectives.

Limits on Generality

We used data from L2014 and L2016 that examined non-Black North American samples on the Race IAT. Future research could investigate whether results are generalizable to other constructs, other implicit measures, and other populations. There are other analytical approaches that researchers could use to assess additional aspects of implicit measure performance such as the ability to override bias (e.g., the Quadruple process model which focusses on accuracy; Conrey et al., 2005; see also Calanchini et al., 2020). We decided to use the diffusion model because it integrates both, accuracy and latency. Thus, it allows computing parameters that are related to meaningful psychological constructs (e.g., participants' speed-accuracy tradeoff; Klauer, 2014).

Conclusion

Our results strengthen prior theorizing by L2014 and L2016, who argued that interventions were effective at temporarily changing associations. Our re-analyses peer into the processes behind these changes and offer insight into *how* interventions impact the decision-making processes underlying categorization (information intake, speed-accuracy tradeoffs, and processes outside of actual decisions). The results add to the current understanding of implicit social cognition.

References

Brandstätter, V., Lengfelder, A., & Gollwitzer, P.M. (2001). Implementation intentions and efficient action initiation. *Journal of Personality and Social Psychology, 81*, 946-960. doi:10.1037//0022-3514.81.5.946

Calanchini, J., Lai, C. K., & Klauer, K. C. (2020). Reducing implicit racial preferences: III. A process-level examination of changes in implicit preferences. *Journal of Personality and Social Psychology,* Advance online publication. https://doi.org/10.1037/pspi0000339

Calanchini, J. & Sherman, J. W. (2013). Implicit attitudes reflect associative, non-associative, and non-attitudinal processes. *Social and Personality Psychology Compass, 7*, 654-667. doi:10.1111/spc3.12053

Cameron, C. D., Brown-Iannuzzi, J. L., & Payne, B. K. (2012). Sequential priming measures of implicit social cognition: A meta-analysis of associations with behavior and explicit attitudes. *Personality and Social Psychology Review, 16*, 330-350. https://doi.org/10.1177/1088868312440047

Conrey, F. R., Sherman, J. W., Gawronski, B., Hugenberg, K., & Groom, C. J. (2005). Separating Multiple Processes in Implicit Social Cognition: The Quad Model of Implicit Task Performance. *Journal of Personality and Social Psychology, 89*, 469-487. doi:10.1037/0022-3514.89.4.469

Devine, P. G. (1989). Stereotypes and prejudice: Their automatic and controlled components. *Journal of Personality and Social Psychology, 56*, 5-18. doi:10.1037/0022-3514.56.1.5

Dovidio, J. F., Kawakami, K., & Gaertner, S. L. (2002). Implicit and explicit prejudice and interracial interaction. *Journal of Personality and Social Psychology, 82*, 62-68. doi:10.1037//0022-3514.82.1.62

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. doi:10.1037//0022-3514.69.6.1013

Fiedler, K., & Bluemke, M. (2005). Faking the IAT: Aided and unaided response control on the Implicit Association Tests. *Basic and Applied Social Psychology, 27*, 307-316. doi:10.1207/s15324834basp2704_3

Finlay, K.A., & Stephan, W.G. (2000). Improving intergroup relations: The effects of empathy on racial attitudes. *Journal of Applied Social Psychology, 30*, 1720-1737. doi:10.1111/j.1559-1816.2000.tb02464.x

Foroni, F., & Mayr, U. (2005). The power of a story: New, automatic, associations from a single reading of a short scenario. *Psychonomic Bulletin & Review, 12*, 139-144. doi:10.3758/BF03196359

Forscher, P. S., Lai, C. K., Axt, J. R., Ebersole, C. R., Herman, M., Devine, P. G., & Nosek, B. A. (2019). A meta-analysis of procedures to change implicit measures. *Journal of Personality and Social Psychology, 117*, 522-559. https://doi.org/10.1037/pspa0000160

Gaertner, S.L., & Dovidio, J.F. (2000). Reducing intergroup bias: The common ingroup identity model. Philadelphia: Psychology Press.

Gaertner, S.L., Dovidio, J.F., Anastasio, P.A., Bachman, B.A., & Rust, M.C. (1993). The common ingroup identity model: Recategorization and the reduction of intergroup bias. *European Review of Social Psychology, 4*, 1-26. doi:10.1080/14792779343000004

Germar, M., Schlemmer, A., Krug, K., Voss, A., & Mojzisch, A. (2014). Social Influence and Perceptual Decision Making: A Diffusion Model Analysis. *Personality and Social Psychology Bulletin, 40*, 217–231. https://doi.org/10.1177/0146167213508985

Gollwitzer, P. M. (1999). Implementation intentions: Strong effects of simple plans. *American Psychologist, 54*, 493-503. doi:10.1037/0003-066X.54.7.493

Green, A. R., Carney, D. R., Pallin, D. J., Ngo, L. H., Raymond, K. L., Iezzoni, L. I., & Banaji, M. R. (2007). Implicit bias among physicians and its prediction of thrombolysis decision for Black and White patients. *Journal of General Internal Medicine, 22*, 1231-1238. doi:10.1007/s11606-007-0258-5

Greenwald, A. G., & Lai, C. K. (2020). Implicit social cognition. *Annual Review of Psychology, 20*, 419-445. doi:10.1146/annurev-psych-010419-050837

Greenwald, A., McGhee, D., & Schwartz, J. (1998). Measuring individual differences in implicit cognition: The Implicit Association Test. *Journal of Personality and Social Psychology, 74*, 1464-1480. doi:10.1037/0022-3514.74.6.1464

Greenwald, A., Nosek, B., & Banaji, M. (2003a). Understanding and using the Implicit Association Test: I. An improved scoring algorithm. *Journal of Personality and Social Psychology*, *85*, 197-216. doi:10.1037/0022-3514.85.2.197

Greenwald, A., Nosek, B., & Banaji, M. (2003b). 'Understanding and using the Implicit Association Test: I. An improved scoring algorithm': Correction to Greenwald et al. (2003). *Journal of Personality and Social Psychology*, *85*, 481. doi:10.1037/h0087889

Haidt, J. (2003). Elevation and the positive psychology of morality. In C. L. M. Keyes & J. Haidt (Eds.), *Flourishing: Positive psychology and the life well-lived* (pp. 275-289). Washington DC: American Psychological Association.

Joy-Gaba, J.A., & Nosek, B.A. (2010). The surprisingly limited malleability of implicit racial evaluations. *Social Psychology, 41*, 137-146. doi:10.1027/1864-9335/a000020

Katz, I. & Hass, R. G. (1988). Racial ambivalence and American value conflict: Correlational and priming status of dual cognitive structures. *Journal of Personality and Social Psychology, 55*, 893-905. doi:10.1037/0022-3514.55.6.893

Klauer, K. C. (2014). Random-walk and diffusion models. In J. Sherman, B. Gawronski, & Y. Trope (Eds), *Dual process theories of the social mind* (pp. 139-152). New York: Guilford Press.

Klauer, K. C., Stahl, C., & Voss, A. (2011). Multinomial models and diffusion models. In K. C. Klauer, A. Voss, & C. Stahl (Eds.), *Cognitive methods in social psychology* (pp. 367-390). New York: Guilford Press.

Klauer, K. C., Voss, A., Schmitz, F. & Teige-Mocigemba, S. (2007). Process components of the Implicit Association Test: A diffusion-model analysis. *Journal of Personality and Social Psychology*, *93*, 353-368. doi:10.1037/0022-3514.93.3.353

Kurdi, B., Seitchik, A.E., Axt, J.R., Carroll, T.J., Karapetyan, A., Kaushik, N., Tomezsko, D., Greenwald, A.G., & Banaji, M.R. (2019). Relationship between the Implicit Association Test and intergroup behavior: A meta-analysis. *American Psychologist, 74*, 569-586. doi:10.1037/amp0000364

Lai, C. K., Hoffman, K. M., & Nosek, B. A. (2013). Reducing implicit prejudice. *Social and Personality Psychology Compass, 7*, 315-330. doi:10.1111/spc3.12023

Lai, C. K., Marini, M., Lehr, S. A., Cerruti, C., Shin, J.-E. L., Joy-Gaba, J. A., . . . Nosek, B. N. (2014). Reducing implicit racial preferences: I. A comparative investigation of 17 interventions. *Journal of Experimental Psychology: General, 143*, 1765-1785. doi:10.1037/a0036260

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*, 1001-1016. doi:10.1037/xge0000179

Lerche, V., & Voss, A. (2017). Retest Reliability of the Parameters of the Ratcliff Diffusion Model. *Psychological Research, 81*, 629-652. doi:10.1007/s00426-016-0770-5

Nosek, B. (2007, April 19). *Implicit Association Test*. [Video file]. Retrieved from https://www.youtube.com/watch?v=n5Q5FQfXZag&feature=related

Nosek, B. A. & Banaji, M. R. (2001). The Go/No-go Association task. *Social Cognition, 19*, 625-666.

Nosek, B. A., Smyth, F. L., Hansen, J. J., Devos, T., Lindner, N. M., Ranganath, K. A., … Banaji, M. R. (2007). Pervasiveness and correlates of implicit attitudes and stereotypes. *European Review of Social Psychology, 18*, 36-88. doi:10.1080/10463280701489053

Olson, M. A., & Fazio, R. H. (2006). Reducing automatically-activated racial prejudice through implicit evaluative conditioning. *Personality and Social Psychology Bulletin, 32*, 421-433. doi:10.1177/0146167205284004

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. doi:10.1037/a0032734

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*, 59-108. doi:10.1037/0033-295X.85.2.59

Ratcliff, R. (2014). Measuring psychometric functions with the diffusion model. *Journal of Experimental Psychology: Human Perception and Performance*, *40*, 870-888. doi:10.1037/a0034954

Ratcliff, R., Smith, P. L., Brown, S., & McKoon, G. (2016). Diffusion decision model: Current issues and history. *Trends in Cognitive Sciences, 20*, 260-281. doi:10.1016/j.tics.2016.01.007

Richeson, J.A., & Nussbaum, R.J. (2004). The impact of multiculturalism versus color-blindness on racial bias. *Journal of Experimental Social Psychology, 40*, 417-423. doi:10.1016/j.jesp.2003.09.002

Riek, B.M., Mania, E.W., & Gaertner, S.L. (2006). Intergroup threat and outgroup attitudes: A meta-analytic review. *Personality and Social Psychology Review, 10*, 336-353. doi:10.1207/s15327957pspr1004_4

Röhner, J., & Ewers, T. (2016a). How to analyze (faked) Implicit Association Test data by applying diffusion model analyses with the fast-dm software: A companion to Röhner & Ewers (2016). *The Quantitative Methods in Psychology, 12,* 220-231. doi:10/20982/tqmp.12.3.p220

Röhner, J., & Ewers, T. (2016b). Trying to separate the wheat from the chaff: Construct- and faking-related variance on the Implicit Association Test (IAT). *Behavior Research Methods*, *48*, 243-258. doi:10.3758/s13428-015-0568-1

Röhner, J., Schröder-Abè, M., & Schütz, A. (2011). Exaggeration is harder than understatement, but practice makes perfect! Faking success in the IAT. *Experimental Psychology, 58*, 464-472. doi:10.1027/1618-3169/a000114

Röhner, J., Schröder-Abé, M., & Schütz, A. (2013). What do fakers actually do to fake the IAT? An investigation of faking strategies under different faking conditions. *Journal of Research in Personality, 47*, 330-338. doi:10.1016/j.jrp.2013.02.009

Röhner, J., & Thoss, P. J. (2018). EZ: An Easy Way to Conduct a More Fine-Grained Analysis of Faked and Nonfaked Implicit Association Test (IAT) Data. *The Quantitative Methods for Psychology, 14*, 17-35. doi:10.20982/tqmp.14.1.p017

Rooth, D. (2010). Automatic associations and discrimination in hiring: Real world evidence. *Labour Economics, 17*, 523-534. doi:10.1016/j.labeco.2009.04.005

Schmitz, F., & Voss, A. (2012). Decomposing task-switching costs with the diffusion model. *Journal of Experimental Psychology: Human Perception and Performance, 38*, 222-250. doi:10.1037/a0026003

Sears, D.O., & Henry, P.J. (2005). Over thirty years later: A contemporary look at symbolic racism. *Advances in Experimental Social Psychology, 37*, 95-150. doi:10.1016/S0065-2601(05)37002-X

Spaniol, J., Madden, D. J., & Voss, A. (2006). A diffusion model analysis of adult age differences in episodic and semantic long-term memory retrieval. Journal of Experimental Psychology*: Learning, Memory, and Cognition, 32*, 101-117. doi:10.1037/0278-7393.32.1.101

Sritharan, R., & Gawronski, B. (2010). Changing implicit and explicit prejudice: Insights from the Associative-Propositional Evaluation Model. *Social Psychology, 41*, 113-123. doi:10.1027/1864-9335/a000017

Strack, F., & Deutsch, R. (2004). Reflective and impulsive determinants of social behavior. *Personality and Social Psychology Review, 8,* 220-247. doi:10.1207/s15327957pspr0803_1

Todd, A. R., & Burgmer, P. (2013). Perspective taking and automatic intergroup evaluation change: Testing an associative self-anchoring account. *Journal of Personality and Social Psychology, 104*, 786-802. https://doi.org/10.1037/a0031999

Turner, R.N., & Crisp, R.J. (2010). Imagining intergroup contact reduces implicit prejudice. *British Journal of Social Psychology, 4*9, 129-142. doi:10.1348/014466609X419901

Van Ravenzwaaij, D., Donkin, C., & Vandekerckhove, J. (2016). The EZ diffusion model provides a powerful test of simple empirical effects. *Psychonomic Bulletin & Review*, *24*, 547-446. doi:10.3758/s13423-016-1081-y

Voss, A., Nagler, M., & Lerche, V. (2013). Diffusion models in experimental psychology: A practical introduction. *Experimental Psychology, 60*, 385-402. doi:10.1027/1618-3169/a000218

Voss, A., Rothermund, K., Gast, A., & Wentura, D. (2013). Cognitive processes in categorical and associative priming: A diffusion model analysis. *Journal of Experimental Psychology: General*, 142, 536-559. doi:10.1037/a0029459

Voss, A., Rothermund, K., & Voss, J. (2004). Interpreting the parameters of the diffusion model: An empirical validation. *Memory & Cognition 32,* 1206-1220. doi:10.3758/BF03196893

Voss, A., & Voss, J. (2008). A fast numerical algorithm for the estimation of diffusion model parameters. *Journal of Mathematical Psychology.* 52, 1-9. doi:10.1016/j.jmp.2007.09.005

Voss, A., Voss, J. , & Klauer, K.C. (2010), Separating response-execution bias from decision bias: Arguments for an additional parameter in Ratcliff's diffusion model. *British Journal of Mathematical and Statistical Psychology, 63*, 539-555. doi:10.1348/000711009X477581

Wagenmakers, E. J., van der Maas, H. L., & Grasman, R.P. (2007). An EZ-diffusion model for response time and accuracy. *Psychonomic Bulletin & Review, 14*, 3-22. doi:10.3758/BF03194023

Table 1 *Computation and Meaning of Diffusion-Model-Based IAT Effects*

Diffusion-model-based IAT effects	Computation	Meaning
IAT_v (ease of decision making)	= parameter *v* incompatible trials - parameter *v* compatible trials	Positive values indicate that it was easier for participants to work on incompatible than on compatible IAT trials, reflecting increased participants' ability to perform the task and increased ease of the task on the incompatible than on the compatible IAT trials. Negative values indicate it was the reverse.
IAT_a (response caution)	= parameter *a* incompatible trials - parameter *a* compatible trials	Positive values indicate that the participants had a more conservative response mode on incompatible than on compatible IAT trials, indicating slow and accurate responding on incompatible trials (here Black-Good and White-Bad) and more speeded and inaccurate responding on compatible trials (here White-Good and Black-Bad). Negative values indicate that it was the reverse.
IAT_{t_0}(non-decision-related processes)	= parameter t_0 incompatible trials - parameter t_0 compatible trials	Positive values indicate that participants took more time for processes outside of the decision process on incompatible than on compatible IAT trials, indicating delays in pressing the response keys and in encoding stimuli on incompatible trials (here Black-Good and White-Bad) than on the compatible trials (here White-Good and Black-Bad). Negative values indicate that it was the reverse.

Figure 1 (a to d). Meta-analytic effectiveness of interventions on overall IAT performance (*D*-score; Fig. 1a), ease of decision-making (IAT_v; Fig. 1b), response caution (IAT_a; Fig. 1c), and non-decision related processes (IAT_{t0}; Fig. 1d). Effective interventions were interventions that changed overall IAT performance in L2014 & L2016. Ineffective interventions were interventions that did not change overall IAT performance in L2014 & L2016. Circles = intervention-level effect sizes; Lines = 95% confidence intervals.