This paper has been accepted for publication at the Journal of Experimental Social Psychology. This is a preprint of the final submitted version. This version of the paper may not be identical to the final published article.

Biases Left Unattended: People Are Surprised at Racial Bias Feedback Until They Pay Attention to Their Biased Reactions

Alexandra Goedderz
University of Cologne, Germany

Adam Hahn
University of Cologne, Germany
University of Bath, UK

Why are people surprised at racial bias feedback, such as test results from Implicit Association Tests (IATs), even though they can predict their IAT racial bias scores prospectively? The present research tested three hypotheses: People are surprised at racial bias feedback due to (1) the feedback wording, (2) implicit evaluations often being preconscious and unattended, or because (3) pretending to be surprised at racial bias feedback is socially desirable. One pilot, four preregistered studies, and a mini-meta-analysis supported hypothesis (2): Although racial biases such as those reflected on IAT scores are observable, people rarely pay attention to them. Specifically, predicting IAT results (Studies 2-4b) and encouragement to pay attention to one's biased reactions before IAT completion (Study 3) reduced surprise, independent of explanation of "implicit bias" (Study 4b). Contradicting the social-desirability hypothesis (3), neither encouragement to admit to bias in the form of abstract predictions (Study 3), nor non-threatening explanations of implicit bias (Study 4b), reduced surprise in the absence of encouragement to pay attention to one's own biases. Speaking against hypothesis (1), surprise was independent of feedback severity (Studies 1-3); and the prediction effect was mediated by recognition of bias, but not correspondence of predictions and feedback (Study 3). These studies suggest that surprise is a consequence of the preconscious nature of automatic social cognitions: People may be motivated to keep consciously accessible racial biases out of awareness. Implications for theories of implicit social cognition and the generality of these effects beyond research on implicit bias are discussed.

Keywords: Attitudes, IAT, implicit bias, preconscious, unconscious

"I was really surprised at the [IAT] results as I never thought of myself as having any biases against Black people." (Study participant).

Recently, the idea that racial biases are widespread even among egalitarian-minded people has increasingly gained traction in public discourse. For instance, Gallup (2021) reports that agreement with the observation that minorities are treated poorly was at an all-time high among Americans of all backgrounds in 2021. One of social psychology's most prominent contributions to this debate – and simultaneously one of its most criticized constructs – has been the concept of implicit bias (BBC News, 2017; CNN, 2015; Scientific American, 2020; The Guardian, 2018, 2021). Implicit bias research has shown wide-spread implicit racial biases across most Western countries among all strata of society (Nosek et al., 2002; Redford, 2018), and in response, implicit bias trainings, in which informing people about their biases is often an important feature, have been on the rise across the world (Chamorro-Premuzic, 2020; Wen, 2020).

In contrast to the observation that acknowledgement of the widespread nature of racial biases is on the rise, however, and in line with the quote in the beginning, research has documented that people respond defensively and with surprise to IAT feedback that communicates that they themselves might harbor racial biases (Howell et al., 2013; Howell et al., 2017; Vitriol & Moskowitz, 2021). Two common explanations for surprise responding to bias feedback have been that racial biases are either purposefully hidden (such that surprise may be a reaction to the disclosure of a hidden response), or that they must be entirely "unconscious" (such that participants could not have known, Haider et al., 2014; Nosek et al., 2002; Quillian, 2008). However, both ideas are at odds with findings that people can predict the patterns of their IAT scores (Hahn et al., 2014) and that people can easily be brought To acknowledge their own (racial) biases (Hahn & Gawronski, 2019). The present research addresses this apparent contradiction: Why are people ostensibly surprised when they learn about their own implicit racial biases (Gawronski, 2019; Howell et al., 2013), even though research suggests that people may be able to predict the patterns of racial bias IAT scores accurately (Hahn et al., 2014)?

Focusing on the IAT as the most widely used measure of implicit bias (Gawronski & De Houwer, 2014), we investigated three explanations: People are surprised because they (1) disagree with the labeling of the racial bias feedback, (2) rarely pay attention to their racial biases, or (3) pretend to be surprised because they believe this is the socially desirable reaction to racial bias feedback. To this end, we gave people feedback to Black-White IATs, measured their surprise reaction, and investigated whether this surprise would decrease in response to (1) different labeling of racial bias feedback, (2) making people pay attention to their racial biases ahead of taking a test, (3a) making people admit to their racial biases, or (3b) describing IAT racial bias scores in more socially acceptable ways. As such, this paper aims to provide evidence about what aspects of IAT racial bias feedback are surprising to lay people and what this can tell us about the – purportedly "unconscious" or "conscious" – nature of the cognitions reflected on implicit bias scores. On a more general level, we suggest that many racial biases – including those reflected on implicit measures such as the IAT – are often "preconscious" (Dehaene et al., 2006; Hahn & Goedderz, 2020): They are rarely attended to, even though, in principle, they are observable. This would not only advance our theoretical understanding of implicit bias, but also point towards simple interventions: It would suggest that people should be encouraged to pay attention to their own reactions to notice their own automatic biases (Hahn & Gawronski, 2019).

Previous Research on Reactions to IAT Racial Bias Feedback

People prefer positive to negative feedback, they want to see themselves in a positive light (Sedikides et al., 2003), and they expect to score better than the average other person on most tests and dimensions (Alicke et al., 1995). From this perspective, it is unsurprising that most people, including people who complete IATs, generally think they are less biased than others (Howell & Ratliff, 2017). Indeed, even as Hahn et al.'s (2014) participants predicted the patterns of their IAT scores accurately, they thought other participants in the same study would show a lot more bias on average – a statistically impossible result. Different from these findings, however, the beginning statement to this paper – if we believe it to be honest – indicates that many people do not just think they are less biased than others, they appear to think that they are not biased at all.

Although there is a thriving research field that investigates defensive reactions to IAT feedback, their causes and consequences, and ways to overcome them (Howell et al., 2013; Howell et al., 2017; Vitriol & Moskowitz, 2021), the specific reaction of surprise has not received similar attention, even though it is often mentioned anecdotally or implied (Gawronski, 2019). For instance, research has found that people are defensive to the degree that their IAT feedback deviates from their explicit evaluations (Howell et al., 2013; Howell et al., 2015; Howell et al., 2017; Howell & Ratliff, 2017). Such defensiveness reactions could indicate that participants are also surprised at being told that they are biased. However, defensiveness and surprise are independent and distinct reactions. For instance, it is possible to be defensive about being told one is biased without being surprised about it (e.g., by expecting a test to be biased); and a person might be surprised without becoming defensive (e.g., new

and unexpected information can be considered interesting). As such, defensive responding to IAT feedback is limited in terms of clarifying whether people are surprised about their bias scores or not. Additionally, surprise is different to defensiveness as it focuses primarily on the feeling of unexpectedness (Stiensmeier-pelster et al., 1995) and may thus be especially fruitful when examining the conscious or unconscious nature of the cognitions reflected on implicit bias scores.

Looking at research that has investigated surprise reactions to IAT feedback specifically, a classroom study by Hillard et al. (2013) showed that more bias feedback on the IAT was associated with higher levels of surprise. However, levels of surprise in this study were rather low (below 2.0 on a 5-point scale with 1 indicating "very slightly or not at all" and 5 indicating "extremely", Hillard et al., 2013, p. 506). Hence, these results question the assumption that people are generally surprised at IAT feedback, but they support the notion that people will tend to be more surprised the more bias the feedback communicates. In a one-item measure Howell et al. (2013) found that participants were more surprised at their IAT feedback the more it deviated from their explicitly reported attitudes. A qualitative analysis by Schlachter and Rolf (2017) looking at comments about IAT feedback on the internet showed somewhat mixed results. Some participants said that they were surprised at their feedback and others that they were not, suggesting that surprise might vary considerably across people and circumstances. In line with this, Perry et al. (2015) found that how their participants reacted to bias feedback was a function of participants' individual differences of bias awareness.

Taken together, there is some evidence that people might be surprised at IAT bias feedback, especially when it deviates from explicit attitudes, but this might differ largely between people. Whether people are surprised at learning that they might harbor any biases (as opposed to no biases) remains an open question awaiting further empirical evidence.

Implicit Evaluations as Unconscious Attitudes

Surprise reactions at IAT feedback are often cited as evidence that the cognitions reflected in implicit measures must be unconscious (Gawronski, 2019; Krickel, 2018; Lane et al., 2007). After all, if people were aware of their biases, they should not be surprised to learn about them. In early debates around implicit bias, the claim that implicit evaluations reflect unconscious attitudes additionally used to often appear in discussions around low correlations between implicit and explicit measures[1] of the targets (Hofmann et al., 2009; Hofmann, Gawronski et al., 2005; Hofmann, Gschwendner et al., 2005; Nosek, 2005, 2007; Nosek & Hansen, 2008). However, low correlations between implicit and explicit measures do not per se speak to the inaccessibility of the cognitions reflected in implicit measures (Gawronski et al., 2006; Hahn et al., 2014; Hahn & Gawronski, 2014). Various prominent dual-process models provide different explanations for why implicit and explicit measures diverge (Fazio, 2007; Gawronski & Bodenhausen, 2006, 2011).

For instance, the MODE model (*M*otivation and *O*pportunity as *DE*terminants) suggests that explicit and implicit evaluations diverge as a function of motivation and opportunity (Fazio, 2007). The main claim here is that people report different evaluations on explicit measures because they are motivated to present themselves in socially desirable ways (Dunton & Fazio, 1997). Gawronski and Bodenhausen's (2011) Associative-Propositional Evaluations (APE) model proposes that people do not always consider the reactions reflected on implicit evaluations to be valid bases for their explicit judgements. According to this model, people are aware of their negative associations with ethnic minorities. However, they might nevertheless report positive attitudes toward them on explicit ratings because they consider other propositional information, e.g. their egalitarian values or specific exemplars of minority members they admire, to be more valid bases for their reported attitudes. As such, both the MODE and APE models argue that the cognitions reflected on implicit measures are generally consciously

[1] We use the term "implicit" to refer to evaluations inferred from indirect computerized reaction time measurements instruments such as the IAT, and "explicit" to refer to self-reported evaluations (Hahn and Gawronski , 2018; De Houwer et al. , 2009). As such, the usage of this terminology makes no assumptions about the underlying cognitions reflected on these measures. We hope to contribute to understanding the underlying cognitions with this paper.

accessible, but often rejected (Fazio & Olson, 2003; Gawronski & Bodenhausen, 2011). And indeed, across several studies Hahn et al. (2014) and Hahn and Gawronski (2019) found that their participants were able to predict the patterns of their implicit evaluations when asked directly. The authors asked their participants to predict how they will score on five different IATs measuring their spontaneous reactions toward five social groups (Black, Asian, Latino, Children, Celebrities) compared to non-celebrity White adults. Their participants were generally good at predicting the patterns of their IAT scores, even though they reported different explicit evaluations. These findings challenge the unconsciousness hypothesis. People can predict the patterns of their implicit evaluations, and there are other explanations for why they report different evaluations when asked explicitly.

Potential Reasons for Surprise

How can the observation that people are surprised at racial bias feedback be reconciled with findings that they can predict the patterns of their IAT scores prospectively? Integrating different theories and empirical evidence with respect to implicit social cognition led us to three different hypotheses.

Surprise and Harshness of Feedback: The Feedback Wording Hypothesis

One hypothesis is that people generally know that they are biased, but they might be surprised at the specific wording that is used to describe their biases (Gawronski, 2019). That is, Hahn et al. (2014) found that participants knew *that* they harbored biases, but they didn't seem to know how biased they were compared to other people; and – consistent with the better-than-average effect (Alicke et al., 1995; Howell & Ratliff, 2017) – they suspected that they were less biased than others.

From this perspective, participants may be surprised at any IAT bias feedback that goes beyond "a slight preference" for one group over another. If this is true, then surprise should be a specific reaction to the (arbitrarily set) conventions and language for IAT feedback, rather than the bias feedback per se (Gawronski, 2019), and people should be less surprised at mild than strong bias feedback. However, as the beginning statement indicates, many people do not only seem to reject the strength of their bias feedback, but the fact that they may harbor any biases at all.

Implicit Evaluations as Preconscious Attitudes: The Attention Hypothesis

If it is in fact true that many people are surprised at harboring any biases at all, then the question remains: How is such surprise compatible with the fact that people can predict the patterns of their IAT scores accurately? Integrating research by Hahn and Gawronski (2019) with theories of consciousness (Dehaene et al., 2006; Hofmann & Wilson, 2010; Hahn & Goedderz, 2020) suggests the following explanation: The cognitions reflected on implicit measures might not generally be unconscious, but often "preconscious" – people rarely pay attention to them unless they are encouraged to do so. In line with the need to view oneself positively, they will hence tend to believe that they are unbiased until they are encouraged to face their biases.

Specifically, Hahn and Gawronski (2019) found that participants aligned their explicit evaluations with their implicit evaluations and acknowledged being biased after they predicted their IAT scores. This indicates that people may learn something new about themselves when they predict their IAT scores. Merely completing IATs (announced as tests of implicit racial attitudes) without predictions changed neither explicit evaluations nor acknowledgment of bias compared to control conditions and pre-test ratings. This last point is important, because it emphasizes that the prediction procedure did not just make participants more honest about cognitions that they knew all along. If that were the case, then knowledge of measurement, and hence completion of IATs, should have had similar effects. Instead, it seems that predicting IAT scores led participants to discover new information about themselves, and this changed their explicit evaluations and their perceptions of how biased they are. Models of consciousness may help clarify this point.

That is, in line with Hofmann and Wilson (2010) and others (Dehaene et al., 2006; Dehaene & Naccache, 2001), we propose that a cognitive process reaches awareness when (1a) it produces a signal that is strong enough, and (1b) attention is paid to this signal. Moreover, a process that produces a detectable signal (1a is present) but

remains outside of conscious awareness because it is left unattended (1b is absent), may be called "preconscious" (Dehaene et al., 2006). A lot of research and theorizing suggest that the signal produced by the cognitions reflected on implicit evaluations is a spontaneous affective reaction (Gawronski & Bodenhausen, 2006; Hahn & Gawronski, 2019; Ranganath et al., 2008; Smith & Nosek, 2011). Integrating these thoughts, people might be surprised at their IAT results because they rarely pay attention to their spontaneous affective reactions to people with different backgrounds. From this perspective, surprise after IAT feedback would indeed be a reaction to learning that one is biased. However, the reason for this surprise is not that the cognitions reflected on implicit measures are generally unconscious. Much rather, surprise would demonstrate that these cognitions are preconscious – people rarely pay attention to them. If this hypothesis is true, then drawing people's attention to their spontaneous affective reactions before receiving IAT feedback should lower their surprise at this feedback.

Real Surprise? The Social Desirability Hypothesis

One last explanation for why people indicate surprise at bias feedback may be social desirability (Crowne & Marlowe, 1960). That is, people may be aware that they harbor biases, but report surprise as an act of self-presentation, because pretending that racial biases are unexpected might be the most desirable answer to give. If this explanation is true and the participant quoted at the beginning of this article was not actually surprised but simply dishonest, then people should always indicate surprise even at "slight" preference feedback- because any level of bias is undesirable. However, this surprise should still be a function of the strength of the feedback. That is, showing "strong" racial preferences is less desirable than showing "slight" preferences, such that reported surprise would have to be a function of the desirability of the specific feedback participants get. Furthermore, presenting the IAT procedure in a non-offensive way should lower participants' surprise because they may perceive their IAT results as less of a threat to their values and beliefs.

Feedback Wording, Attention, or Social Desirability?

To test these three hypotheses, the four studies presented in this paper measured surprise reactions in response to IAT racial bias feedback. Although the three hypotheses are compatible in some instances, we designed our studies such that they would answer three empirical questions to which the three hypotheses make opposing predictions. The first is whether or not participants report more surprise in response to all levels of racial bias feedback - even low levels of bias - when compared to no-bias feedback. The second is whether this surprise is a function of the degree of bias the feedback communicates, such that feedback of a "strong" bias would lead to more surprise than feedback of "mild" bias. The last is whether making people pay attention to their biased reactions before test completion reduces surprise. The three hypotheses' predicted answers to the three questions are summarized in Table 1 and described next.

The feedback wording hypothesis predicts no surprise at bias feedback that is clearly at the low end of the scale, but increased surprise the more comparative bias the feedback indicates. This surprise reaction should further not change when people are asked to pay attention to their biases before IAT completion. According to the feedback wording hypothesis, participants already know that they harbor biases, and hence asking them to pay attention to those biases should not lead to any new insights.

The attention hypothesis predicts surprise at any feedback indicating bias unless the person is encouraged to first pay attention to their biased reactions. The attention hypothesis makes no predictions regarding reactions to severity of feedback.

Lastly, the social desirability hypothesis predicts both surprise at all bias feedback and increased surprise the less socially desirable said feedback sounds. Making participants pay attention to their biases before IAT completion may reduce "pretend surprise", but only to the degree that it either induces participants to admit to biases they would otherwise hide, or changes their perception of what a socially desirable response is. Hence, it should not require any specific *attention*-to-bias manipulation.

Table 1
Empirical Questions in the Present Research and Their Predicted Outcomes According to the Three Hypotheses

		Hypotheses		
		Feedback wording hypothesis	Attention hypothesis	Social-desirability hypothesis
Reason for surprise at IAT feedback		Arbitrary labels: People know their biases but disagree with the labels.	Preconscious attitudes: People rarely pay attention to their biases.	Pretend surprise: People know their biases but admitting this is undesirable.
Empirical Questions	1. Do people generally report surprise at any racial bias feedback compared to no-bias feedback, including low levels of bias?	No	Yes	Yes
	2. Is the level of surprise a function of the degree of bias communicated in the feedback?	Yes	Both answers compatible/no prediction	Yes
	3. Does paying attention to one's biases before IAT completion reduce surprise at IAT feedback?	No	Yes	No (but see text)

Instead, any request to admit to biases before IAT completion should suffice to both induce a person into admitting bias and shift their perception of a desirable response. We explain these last points in more detail in Study 3 when we test them directly.

In sum, observing the patterns of results to the three questions we investigated allowed us to see which hypothesis explains best why people report surprise at IAT bias feedback.

The Present Research

The aim of the present studies was to investigate three potential explanations for why people react with surprise at racial bias feedback even though they can predict the patterns of their IAT scores prospectively: (1) disagreement with the feedback wording, (2) the preconscious nature of implicit attitudes (attention hypothesis), or (3) pretend surprise due to social desirability concerns.

We started this research project with a pilot study in which we asked participants to imagine hypothetical feedback to test the surprise scale we developed for subsequent studies. In Study 1, we tested whether people in fact react with surprise to performance-based bias feedback compared to feedback that declared "no meaningful bias" (Question 1, see Table 1). To test the feedback wording and the social desirability hypotheses, Studies 1-4 further investigated whether surprise was a function of the degree of bias communicated in the feedback (Question 2, see Table 1). Additionally, we altered the feedback to be less socially undesirable in Study 2. Addressing the attention hypothesis, Studies 2-4 tested whether participants would be less surprised at IAT feedback after encouragement to pay attention to their spontaneous affective reactions to stimuli of the targets in question (Question 3, see Table 1).

We first operationalized attention to reactions as predicting IAT scores before completing IATs (Studies 2-4, Hahn & Gawronski, 2019). In Studies 3 and 4b, we then disentangled whether the effect of predictions on surprise could be better explained by attention, social desirability concerns, and/or the wording of the feedback. Specifically, Study 3 tested whether people simply pretend to be biased unless they are induced to admit to biases ahead of time (social desirability hypothesis), or if they are instead truly surprised at IAT feedback as long as they are not encouraged to

pay attention to their biased reactions (attention hypothesis). Study 4b investigated whether non-threatening information about implicit evaluations and the IAT could explain the prediction effect in the absence of attention to one's affective reactions (social desirability hypothesis). Finally, to compare the attention and feedback wording hypotheses directly, a mediation analysis in Study 3 also tested whether the effect of prediction on surprise could be better explained by correspondence of feedback with expectations (feedback wording hypothesis) or by recognition of bias (attention hypothesis).

We preregistered all studies (except for the pilot study) and report all data, measures, manipulations, and exclusions in each study to allow for increased transparency, replicability, and trustworthiness of our findings (Lindsay et al., 2016). Data analyses were only conducted once the full samples reported here were collected and preregistered data exclusions were completed. We report all preregistered analyses and indicate where we conducted non-preregistered analyses. All materials, data sets, preregistrations, and analysis files can be found at https://osf.io/bezkx/.[2]

Pilot Study

This study was aimed at piloting a scale developed to measure surprise at IAT feedback in the present line of research. The study also tested whether people would be more surprised when imagining bias as opposed to no-bias feedback. It was not preregistered.

Method

Participants

One-hundred and twenty-two participants were recruited on Amazon's Mechanical Turk platform (MTurk) in exchange for US-$ 0.10 basic payment and US-$ 0.10 possible bonus payment. After excluding eight participants who failed at least one of two attention check items, the final sample consisted of 114 participants [3] (52.6% female; median age = 36, age range = 18-64 years). All participants were American citizens and most (76.3%) identified as White (8,8% Black/African-American, 7,9% (East-) Asian, 0.9% Latino/Hispanic, 6.1% more than one ethnic category).

Materials and Procedure

Participants were asked to imagine they were to receive feedback on a Black-White IAT that either reveals that they have *"...a strong automatic preference for WHITE over BLACK*" (strong-bias condition) or "…*NO statistically detectable preference for either BLACK or WHITE*" (no-bias condition), as well as to repeat this feedback on the next page (attention check). Seeing their feedback on top of the screen, participants then completed a ten-item surprise scale aimed at self-reported surprise and unexpectedness of IAT feedback (see Table 2 for final six items and their psychometric properties, and OSF repository for all ten initial items), on 7-point Likert scales ranging from 1 (*strongly disagree)* to 7 (*strongly agree*), and in individually randomized orders.

The scale included another attention check item ("Attention-Check: Please select '3'"). The study concluded on demographic information and a debriefing.

Results and Discussion

A *t*-test on the average of the ten items of the surprise scale showed that participants who imagined receiving strong-bias feedback scored higher ($M = 4.94$, $SD = 1.76$) than participants who imagined getting no-bias feedback ($M = 2.39$, $SD = 1.60$), $t(112) = 8.11$, $p < .001$, $d = 1.49$, 95% CI [1.08, 1.91]. This result also replicated on every item individually (all ts > 6.83, all ps < .001)[4], such

[2] We chose the preregistration template from "as predicted", which is not stored by name on the Open Science Framework (OSF). Hence, the preregistrations can only be identified by their dates within the project, which we indicate in each study. We also provide direct links to each registration.

[3] We conducted power sensitivity analyses using G*Power (Faul et al., 2007) for all studies. With the final sample size of 114 (53 participants in the bias condition, and 61 in the no-bias condition; assuming alpha = 0.05; two-tailed; power = 80%), this pilot study had to have a minimum effect size of $d = 0.53$ to show up as significant, which could be reached at a critical t value of 1.98.

[4] There were no noticeable differences in psychometric quality between the 10 items. All ten items loaded on the same first component in a principal components analysis (PCA, 89.8% of variance explained) and the original ten-item reliability (Cronbach's $\alpha = .99$) never dropped below .98 from the exclusion of any particular item.

that we selected a smaller number of three forward and three backward-phrased items for the remaining studies (see Table 2, Cronbach's α = .98).

Table 2

Item Factor Loadings on The First Component of a Principal Components Analysis, and Variance Explained by This Component, across all 4 studies and the pilot.

		Factor Loadings per Study				
	Item	Pilot	1	2	3	4b
1.	I was surprised at my IAT result.	.96	.91	.88	.89	.85
2.	I expected a different IAT result.	.97	.91	.90	.87	.87
3.	I did not expect to get the IAT result that I got.	.96	.89	.92	.92	.88
4.	My IAT result confirmed what I expected to find. (rev.)	.98	.95	.92	.94	.89
5.	My IAT result supported my initial expectation. (rev.)	.88	.94	.91	.93	.90
6.	I expected the IAT to show the result that it showed. (rev.)	.95	.95	.94	.91	.90
% of Variance explained		89.78	85.83	83.04	82.79	78.03
Cronbach's alpha		.98	.97	.96	.96	.94

Taken together, the pilot established our surprise scale and provided first evidence for Question 1: participants reacted with more surprise when they imagined receiving feedback of harboring strong racial biases as opposed to feedback that they harbor no biases.

Study 1

Study 1 aimed at providing empirical evidence concerning whether people are surprised at IAT feedback indicating racial biases.[5] Specifically, we tested whether surprise at standard feedback as typically given at www.projectimplicit.com would be higher than performance-independent no-bias feedback (Question 1, see Table 1), as well as whether surprise would be a function of the degree of bias the feedback indicates (Question 2, see Table 1). The preregistration for this study can be found at https://osf.io/8uev5/ and was registered on January 10th, 2019.

Method

Participants and Design

Study 1 featured a two-condition between-subjects design (standard feedback vs. no-bias feedback). A G*Power analysis (Faul et al., 2007) for an independent samples *t*-test with an allocation ratio of two-to-one (allocating twice as many participants to the standard feedback condition as to the no-bias feedback condition) revealed that, to find a medium sized effect of $d = .50$ with 80% Power, we would need at least 144 participants. To account for our preregistered exclusion criteria, we aimed at recruiting 180 participants via the service TurkPrime on Amazon's Mechanical Turk (MTurk). Participants were first informed that the study contained a computerized reaction time task, and that they could only proceed to the end if they followed instructions and would not click buttons randomly on this task. Participants who responded with <= 300ms in 10% or more of the trials were excluded from the study after IAT completion (Greenwald et al., 2003). All participants who completed the study received a basic payment of US $0.70. Another US-$0.30 were paid if the two attention check items also used in the pilot were answered correctly. In total, 198 participants started the study on MTurk, of which 180 completed all tasks and passed the preregistered

[5] We report the second attempt at running the exact same study. Our first study revealed a minor programming mistake. When participants completed the surprise scale in that study, the title instructions encouraged them to "imagine" they received feedback on an IAT, even though they had just in fact received feedback on IATs, so we decided to run the study again. The main effect largely remained the same so that we decided to only report the second study (see supplemental materials). The preregistration for the first study can be found at https://osf.io/rnf3j/ and was registered on December 12th, 2018. All materials, data, and analyses are available online.

exclusion criteria [6,7] (50% Female; median age = 33, age range = 18-65 years, 98.9% US-American citizens). 72.8% of the participants self-identified as White (7.8% Black/African American, 3.3% Latino/Hispanic, 6.1% East-Asian, 3.9% South-Asian, 6.2% more than one ethnic category or another ethnicity).

Black-White IAT

To create a seven-block IAT (Greenwald et al., 1998) in Qualtrics, we used the IATgen tool (https://iatgen.wordpress.com, Carpenter et al., 2019). As targets, we used pictures of 10 Black and 10 White individuals (five male and five female individuals per target category) adapted from Hahn et al. (2014) who used pictures from the productive aging lab database (Minear & Park, 2004). The attributes consisted of 10 positive and 10 negative words (see all stimuli on the OSF repository at https://osf.io/yxrpb/). In the first block, participants completed 20 practice trials categorizing the attribute words to the left or right side by using the E or I key on their computer keyboards. The second block consisted of another 20 practice trials categorizing the target pictures to the left or right side as "White" or "Black". Blocks 3 (20 trials) and 4 (40 trials) were combined blocks where participants had to either react with one key to pictures of Black people and negative words, and with the other key to White people and positive words (prejudice-compatible), or the other way around (prejudice-incompatible). In the fifth block, the target pictures switched sides and participants spent 40 trials practicing the reversed categorization. Block 6 and 7 were structurally similar to Blocks 3 and 4, but with a changed combination. Participants who completed the prejudice-compatible blocks first now completed the prejudice-incompatible blocks and those who first completed the prejudice-incompatible blocks now completed the prejudice-compatible blocks. The order of blocks (prejudice-compatible or prejudice-incompatible first) as well as the key-assignments (good-left, bad-right or bad-right, good-left) were randomly assigned between subjects. As such, participants completed one of four possible IAT combinations. When participants made an error, they were shown a red "X" and asked to correct their response by pressing the other button. Reaction times were measured from the stimulus onset until participants indicated the correct response (Greenwald et al., 2003). A *D*-Score was computed according to Greenwald et al. (2003), dividing the reaction time differences for Block 3 and 6 (incompatible – compatible) by the pooled standard deviation of both blocks. The same was done for Blocks 4 and 7. The final *D*-Score was derived from the mean of these two scores with a positive value indicating faster reaction times in the compatible blocks compared to the incompatible blocks, which is interpreted as a pro-White bias. Negative scores indicate a pro-Black bias. The IAT showed satisfactory reliability (Cronbach's alpha = .77, calculated from the two *D*-scores).

IAT Feedback

The original Javascript Code by Carpenter et al. (2019) was altered to calculate a *D*-Score within Qualtrics that was used to present a feedback statement to participants: "*Your data suggest […] automatic preference for [Group A] over [Group B]"*. Two-thirds of the participants received a feedback statement with qualifiers based on conventions used on http://www.implicit.harvard.edu: *little to no* for $|D| <= .15$, *a slight* for $.15 < |D| <= .35$, *a moderate* for $.35 < |D| <= .65$, and *a strong* for $|D| >= .65$. The groups were imputed depending on the sign of the *D*-Score (i.e., "…preference for WHITE" or "…BLACK"). The remaining third of the participants received performance-independent feedback: "*Your data suggest NO meaningful automatic preference for either BLACK or WHITE".*

Procedure

All participants first provided informed consent and were informed about possible bonus payments and exclusions. Afterwards, participants

[6] With the final sample size of 180 (56 participants in the no-bias feedback condition, and 126 in the standard-feedback condition; assuming alpha = 0.05; two-tailed; power = 80%), Study 1 had to have a minimum effect size of $d = 0.46$ to show a significant effect, which could be reached at a critical t value of 1.97.

[7] As preregistered, although 25 participants failed to answer the attention check regarding their feedback correctly, they were kept in the final sample. In the previous study, 20 participants were excluded due to this exclusion criterion. We decided to not honor this exclusion criterion again because (1) it excluded participants unequally from conditions, (2) results remained largely the same with or without these participants, and (3) participants were reminded of their actual IAT feedback before IAT completion such that it was ensured without this attention check that they knew their IAT result.

were randomly assigned to either the "standard feedback" (2/3rd of participants) or the "no-bias feedback" (1/3rd of participants) condition. In both conditions, participants received a brief introduction to the Black-White IAT, were told that they would receive feedback on their IAT, and then completed the IAT. Next, all participants received their respective feedback, followed by an attention check item that asked them to indicate which of several feedback options they had just received. Finally, all participants completed the surprise scale (Cronbach's alpha = .97, see Table 2), demographic information, and were given the chance to comment on the study. At the end, all participants were debriefed about the purpose of the study, including a detailed explanation pertaining to feedback scoring conventions of the IAT.

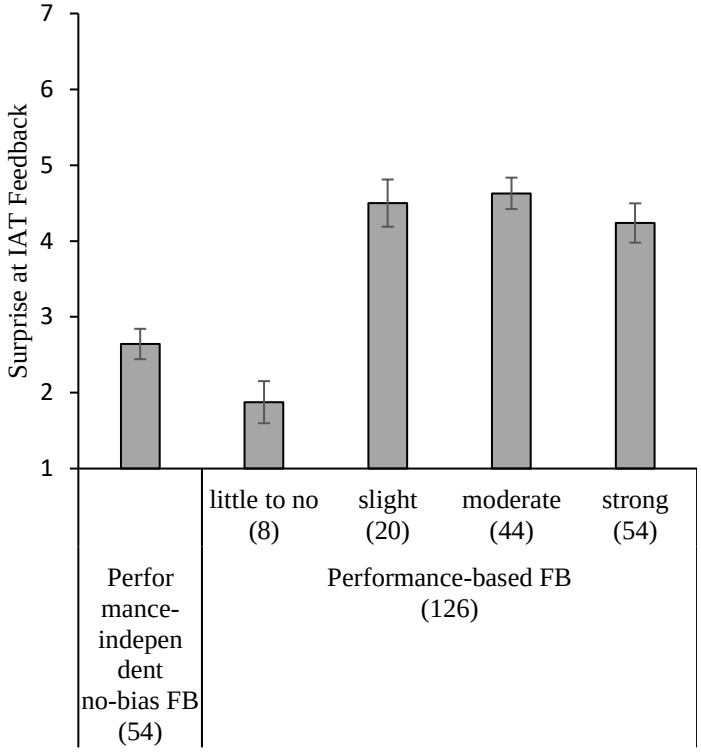
Results

An independent-samples *t*-test on the average of the six surprise items showed that participants who received standard IAT feedback (M = 4.27, SD = 1.71) were more surprised than those who received no-bias feedback independent of performance (M = 2.64, SD = 1.47), $t(178) = 6.07$, $p < .001$, $d = 0.99$, 95% CI [0.65, 1.32].

To investigate whether participants were more surprised the stronger the wording of their bias feedback (Question 2, see Table 1), we looked at the correlation between surprise and the absolute feedback on the IAT ("Slight" = 2, "Moderate" = 3, "Strong" = 4) for the group who received bias feedback based on their performance. As explained in Table 1, we consider the effects of the strength of the wording of the feedback on surprise (i.e., Question 2) a different question than whether *any* bias feedback leads to more surprise than no-bias feedback (Question 1). Hence, participants who received "little to no" bias feedback were excluded from this analysis on the effects of feedback wording (See Figure 1). There was no significant correlation between surprise and the feedback categories, $r(118) = -.08$, $p = .374$, and neither were there any other significant differences in surprise between the three bias feedback conditions, all ps > .24 (see no rise in surprise between the bars that indicate "slight", "moderate", or "strong" bias in Figure 1).

Figure 1

Study 1: Level of Surprise as a Function of IAT Feedback (FB) on Racial Preferences



Note. Error bars represent population-estimated standard errors. Numbers in parentheses represent number of participants who received said feedback.

Discussion

The results of Study 1 confirmed the so-far anecdotal observation that people react with surprise to IAT feedback indicating bias as opposed to no-bias feedback, independent of the severity and wording of the feedback. Even participants who were told that they had a "slight preference" were more surprised than participants who were told that they were not biased, but no less surprised than participants who were told they had a "strong preference".

These findings are in line with the attention hypothesis, which states that, when not encouraged to pay attention to their own biases, people's need to see themselves positively leads them to believe and expect to be unbiased, independent of the severity of the feedback. At the same time, they provide first evidence against the feedback wording hypothesis, which would not predict surprise at low-bias feedback, and increased

surprise the more severe the feedback. Although the social-desirability hypothesis predicts that participants will pretend to be surprised at any IAT bias feedback including low-bias feedback, it would also predict a relationship between harshness (= undesirability) of feedback and surprise (see Table 1). Hence, the results of Study 1 so far favor the attention hypothesis as an explanation for surprise reactions to IAT bias feedback.

Study 2

The aim of Study 2 was to experimentally test whether the attention or the feedback wording hypothesis is better suited to explain why people report surprise at IAT bias feedback (as Study 1 showed). To manipulate the attention people pay to their spontaneous biased reactions, participants either predicted their IAT scores or not. Hahn and Gawronski (2019) found that people tend to discover previously unattended biases when they predict IATs. Building on these findings, we reasoned that, if the attention hypothesis is true, participants who complete predictions should be less surprised at their IAT results than participants who do not complete predictions.

To test the feedback wording hypothesis, which states that the strength of bias communicated in the feedback predicts the level of surprise people report, we either gave participants standard IAT feedback as used on www.implicit.harvard.edu, or they received reduced feedback; thus complementing the correlational findings of Study 1 with an experimental manipulation. The reduced feedback simply said that the IAT indicated "an automatic preference" for one group over the other, without adding qualifications based on the degree of bias. If the feedback wording hypothesis is true, we reasoned, then participants who receive reduced feedback should be less surprised at their IAT feedback than those who receive standard feedback, independent of whether they predicted their IAT scores or not. The preregistration for this study can be found at https://osf.io/fh542/ and was registered on December 18th, 2018.

Method

Participants and Design

The study featured a 2 (prediction vs. no prediction) by 2 (standard feedback vs. reduced feedback) between-subjects design. A power analysis using G*Power (Faul et al., 2007) indicated that to find a small to medium effect size $f = 0.20$ with at least 90% power we would need a total sample size of $N = 265$. We rounded this number and aimed at recruiting 300 participants via TurkPrime. Participants who completed all parts of the study received a basic payment of US$ 0.80 for participation and another US$ 0.40 if they correctly answered two attention check items embedded in the study. Overall, 349 participants started the study of which 28 instantly opted out. As preregistered, 17 participants who responded too fast on the IAT (Greenwald et al., 2003) were dropped from the study, and another two participants were excluded from data analysis because they failed the attention check items. The final dataset consisted of 302 participants[8] (55.6% Female; median age = 32, age range = 19-72 years). 97.4% of the participants reported being US-American citizens and 91.7% were born in the USA. 70.5% self-identified as White/Caucasian (9.9% Black/African American, 5.6% Latino/Hispanic, 5.0% East-Asian, 2.6% South-Asian, 6.3% more than one of the ethnic categories or another ethnicity).

Materials

Prediction Task. Participants in the prediction condition first received an introductory text explaining the concept of implicit attitudes and the IAT as "spontaneous affective reactions" that often differ from what people would express when asked directly. Then they completed a trial prediction towards cats and dogs before they completed the actual prediction towards Black and White people. In it, they were asked to look at the pictures that represent the social categories Black and White (that were also used in the IAT), and to listen to their gut reactions to predict what an IAT on these social categories would show. The prediction question said "I predict that the IAT comparing my reactions to BLACK vs WHITE will show that my implicit attitude is…" with a scale

[8] With the final sample size of $N = 302$ (randomly assigned to one of four conditions; assuming alpha = 0.05; power = 80%), Study 2 had to reach a minimum effect size of $f = 0.16$ to show a significant effect, which could be reached at a critical F value of 3.87.

ranging from -3 ("…a lot more positive toward BLACK") to 3 ("…a lot more positive toward WHITE"). Participants in the no-prediction condition completed four similarly-formatted filler items on consumer preferences (casual vs. formal cloths, junk food vs. vegetables, texting vs. talking on cellphone, outdoor vs. indoor activities), but the topic of race was not mentioned.

IAT Feedback. All participants received feedback based on their performance on the IAT. While participants in the standard-feedback condition received feedback according to the same conventions as described in Study 1, participants in the reduced-feedback condition were only told that they showed either "little to no automatic preference" (|D| < .15) or "an automatic preference" (|D| >=.15).

Procedure

Participants were first informed about all conditions of payment and participation, including possible bonus payments and exclusions, provided standard informed consent, and were randomly assigned to one of the four conditions. Participants then completed the predictions as explained above, or the filler items. Afterwards, all participants read the introduction to the IAT informing them that they will receive feedback, completed the same Black-White IAT as described in Study 1 (Cronbach's alpha = .73), and received feedback on their performance on the IAT depending on condition. Next, participants were asked to recall and indicate the feedback they received to increase attention to it. Finally, all participants completed the surprise scale (Cronbach's alpha = .96, see Table 1) with their feedback repeated on top. They ended the study with demographic information, a chance to comment on the study, and finally, a debriefing. This debriefing included an explanation and discussion on the different feedback conditions, similar to Study 1.

Results

Level of Bias

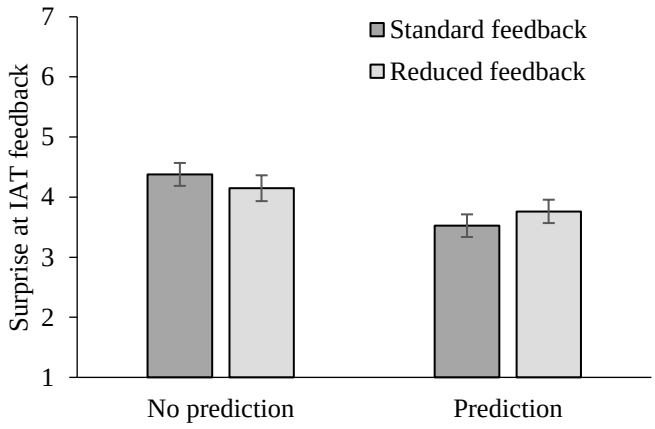
To explore whether the prediction manipulation prior to completing the IAT influenced participants' bias scores, we ran an independent-samples *t*-test on average *D*-scores (not preregistered). There was no statistically significant difference in D-scores between participants who predicted their IAT results (M = 0.47, SD = 0.40) and participants who did not (M = 0.54, SD = 0.39), $t(300) = 1.66$, $p = .098$, $d = 0.19$, 95% CI[-0.04; 0.42].

Surprise at IAT Feedback

We conducted a 2 (prediction vs. no prediction) x 2 (standard feedback vs. reduced feedback) between-subjects ANOVA on average responses on the surprise scale. Results showed a significant main effect of prediction, $F(1, 298) = 9.83$, $p = .002$, $\eta_p^2 = .032$, indicating that participants who completed predictions were less surprised at their IAT feedback than those who did not complete predictions (see Figure 2).

Figure 2

Study 2: Level of Surprise by Experimental Condition



Note. Level of surprise as a function of IAT score prediction and type of IAT feedback received. Errors bars depict standard errors of estimated marginal means from a 2 (IAT score prediction vs. no prediction) x 2 (standard IAT feedback vs. reduced IAT feedback) ANOVA. N = 302, randomly assigned to condition.

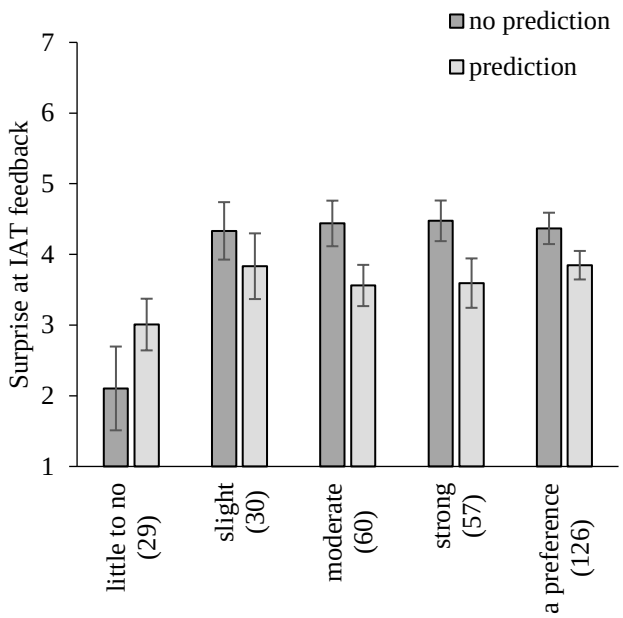
A non-significant main effect of the type of feedback, $F(1, 298) < 0.01$, $p = .983$, $\eta_p^2 < .001$, further showed that there was no significant difference in surprise between participants who received feedback with potentially threatening and undesirable labels compared to bias feedback without specific qualifiers. Additionally, there was no significant interaction, $F(1, 298) = 1.39$, $p = .239$, $\eta_p^2 = .005$, confirming that independent of the type of feedback, completing predictions made participants less surprised at their IAT feedback.

Surprise and Strength of Bias Feedback

We again looked at the relationship between surprise and IAT feedback labels (see Study 1), including only participants who received standard feedback and excluding participants who received "little to no" bias feedback.

Figure 3

Study 2: Level of Surprise as a Function of IAT Score Prediction for the Five Different Performance-based Feedback Labels



Note. Level of surprise as a function of IAT score prediction for the five different performance-based feedback labels participants received. Errors bars depict standard errors of estimated marginal means from a 2 (IAT score prediction vs. no prediction) x 5 (levels of feedback) ANOVA. Numbers in parenthesis represent number of participants who received said feedback.

As can be seen in Figure 3, degree of surprise was again independent of the level of bias indicated by the feedback, $r(147) = .01$, $p = .896$ (see no rise in surprise for the three pairs of bars in the center of the figure). The same result emerged when we included participants in the reduced feedback condition who were told that they have "an automatic preference" (= 2; equivalent to the "slight" feedback category), $r(273) < -.01$, $p = .980$. Hence, this analysis continued to find no evidence that the strength of the bias communicated in the wording of the feedback was the reason for the surprise reactions (Gawronski, 2019).

Prediction Accuracy

Although not the primary purpose of the present study, we decided to examine the correlation between IAT score predictions and actual IAT scores. Hahn et al. (2014, see also Hahn & Goedderz, 2020) have argued that awareness of the cognitions reflected on IAT scores can only be analyzed in within-subjects correlations across several IATs and several predictions. This is because between-subjects correlations between predictions of one IAT and IAT scores require that people calibrate their attitudes consistently: A more biased person needs to use a stronger bias label than a less-biased person. Notwithstanding these limitations (we only administered one IAT, precluding within-subjects analyses), the between-subjects prediction accuracy in the current sample was significant, $r(159) = .42$, $p < .001$, indicating both awareness of bias and consistent social calibration across participants.

Surprise and Deviation of Predictions From Feedback

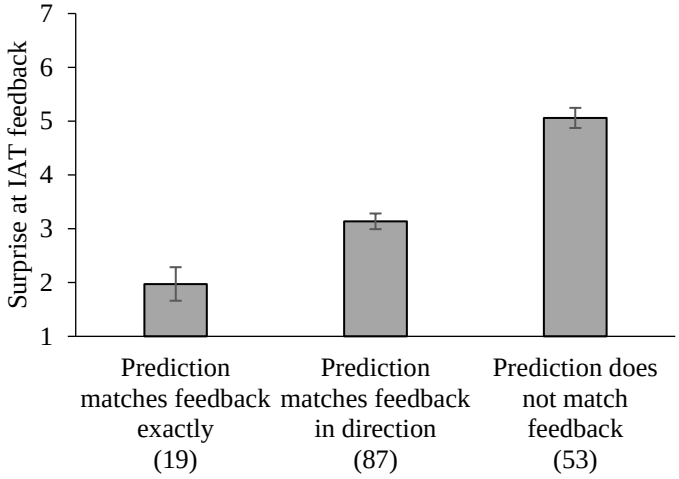
As a last step, we conducted several non-preregistered analyses, investigating whether participants' reported surprise was a function of the deviation between their predictions and feedback. To this end, we coded all responses in the standard feedback condition to create a variable with three levels. They indicated whether participants (1) predicted IAT scores that exactly matched their specific IAT feedback (e.g., predicting a "slight preference for WHITE over BLACK" and receiving feedback indicating "a slight preference for WHITE over BLACK", 22.0% of participants); (2) predicted their IAT scores to be in the same *direction* as the feedback (e.g., correctly predicting a preference for one group over the other, independent of the degree of bias, 51.2%); or (3) did not predict the same direction of bias as the feedback (e.g., predicting no preference, but receiving feedback of a preference for one group over the other, 26.8%). For participants in the reduced feedback condition, 58.4% predicted the same direction of bias as the feedback, and 40.3% predicted their scores to be opposite to the IAT feedback or predicted no bias. One participant predicted showing "little to no bias" and actually received that feedback.

As can be seen in Figure 4, participants' level of surprise significantly differed between the

three categories, $F(2, 156) = 49.11$, $p < .001$, $\eta_p^2 = .386$.

Figure 4

Study 2: Level of Surprise by Correspondence of Predictions and Feedback for Participants in the Prediction Conditions



Note. Level of surprise as a function of correspondence between predictions and feedback for participants in the prediction conditions (N = 159). Error bars depict standard errors of estimated marginal means from a one-way ANOVA testing differences between the three conditions. Numbers in parenthesis represent number of participants in each category.

Examining contrasts further revealed that participants who received feedback that exactly matched their predictions were less surprised (M = 1.97, SD = 0.89) than those who predicted their IAT scores with different labels (M = 3.14, SD = 1.39), $F(1, 156) = 40.58$, $p < .001$, $\eta_p^2 = .206$. However, those participants whose feedback indicated bias in the same direction as they predicted were still less surprised than the midpoint of the scale (4), $t(86) = 5.77$, $p < .001$, and than those who received feedback that was entirely inconsistent with their predictions (M = 5.06, SD = 1.44), $F(1, 156) = 97.21$, $p < .001$, $\eta_p^2 = .384$.

Consistent with Howell et al. (2013), we also found a strong relation between surprise and the absolute deviation of participants' feedback from their predictions, $r(159) = .55$, $p < .001$. Hence, in contrast to the analyses on continuous feedback, these analyses suggest that feedback wording does contribute to surprise at IAT feedback, at least to the degree that it deviates from participants' stated expectation (Gawronski, 2019; Howell et al., 2013).

Discussion

Study 2 tested whether attention or the strength of the feedback wording could explain why people react with surprise to IAT feedback indicating bias. The attention hypothesis states that people react with surprise because they rarely pay attention to their biases. The feedback wording hypothesis states that people react with surprise at the wording used to describe their biases. The data supported the attention hypothesis. Participants who were asked to pay attention to their spontaneous affective reactions by predicting their IAT results were less surprised at their IAT feedback than participants who did not predict their IAT results.

Concerning the feedback wording hypothesis, results were mixed. Participants were equally surprised when they received standard feedback as when they received reduced feedback, and strength of feedback was again uncorrelated with surprise. At the same time, people were more surprised the more their feedback deviated from their predictions. One interpretation of these findings is that, even though participants often choose different labels for their biases, it is not the *harshness* of the feedback that drives this effect. Another interpretation is that omitting all qualifiers in the feedback was not perceived as less harsh, such that our manipulation did not work as intended. We will return to this point in the general discussion.

Although the effect of IAT score predictions on surprise is a direct prediction from the attention hypothesis, it is also compatible with the social-desirability hypothesis if we assume that surprise reports are dishonest. Prediction might lower people's surprise not because it made them pay attention to their biases, but because it made them admit to biases ahead of time. This would make a pretend-surprise reaction superfluous, and it may shift perception such that admitting to biases would now become a socially desirable response. Additionally, the prediction manipulation may have presented the IAT in more socially desirable ways, such that the feedback became less threatening. Studies 3 and 4 aimed at investigating these alternative explanations.

Study 3

The purpose of Study 3 was to investigate whether the effect of IAT score prediction on surprise could be explained by attention to biased reactions, or by social desirability concerns. Specifically, the prediction procedure consists of three steps (see Table 3). Relevant for this study, it includes a manipulation to pay attention to one's biased reactions (explained in Step 1, executed in Step 2), and a request to predict these biased reactions on 1-7 scales (Step 3). The attention hypothesis states that Steps 1 and 2, the process of paying attention to one's biased reaction, is responsible for the reduced surprise reactions. However, according to the social desirability hypothesis, Step 3, the overt rating of this reaction on a scale, may be responsible for the effect. Specifically, concrete predictions may induce participants to "admit" to biases of which they are always fully aware, but which they would otherwise hide. It might also shift participants' perception such that admitting to biases becomes the more desirable response. If this hypothesis is true, we reasoned, then any question that induces participants to admit to harboring biases ahead of time should reduce reported surprise, even if it does not include any encouragement to pay attention to one's biased reactions.

To test these competing explanations, we designed a 2-by-2 design in which we independently manipulated whether participants were asked to predict their results on the prediction scale or not (see Table 3, Step 3, prediction manipulation), and whether they were asked to pay attention to their spontaneous affective reactions or not (see Table 3, Steps 1 and 2, attention manipulation). If the completion of the prediction scale reduced surprise in Study 2 because it induced participants to admit to biases they knew all along and thus shifted perceptions of social desirability, then this should result in a main effect of the prediction manipulation in this study. On the other hand, if the effect is driven by attention to one's spontaneous reactions, then this should result in a main effect of the attention manipulation. The preregistration can be read at https://osf.io/ze3pg/ and was registered on February 21st, 2019.

Table 3

Components of the Prediction Procedure Included in Each Condition in Study 3. The Exact Wording and All Materials Can Be Found on OSF.

		Conditions			
		Attention		No Attention	
		Prediction	No Prediction	Prediction	No Prediction
		Standard prediction			Control
	Prediction procedure				
Attention Manipulation	Step 1: Explanation of IAT as measure of spontaneous affective reactions Step 2: Pay attention to reactions to pictures used in the IAT	Yes	Yes	-	-
Prediction Manipulation	Step 3: Completion of prediction scale	Yes	-	Yes	-

Note. The explanation in Step 1 included a non-threatening explanation of implicit attitudes as spontaneous reactions that may be different from explicitly endorsed attitudes. The no-explanation version simply introduced the IAT as a test of "implicit attitudes" that would reveal a preference for BLACK or WHITE. The control condition included filler items about consumer preferences without mentioning race or bias. Steps 2 and 3 were always completed twice, once hypothetically towards cats and dogs, and then towards BLACK and WHITE

Method

Participants and Design

Study 3 consisted of a 2 (attention vs. no attention) by 2 (prediction vs. no prediction) between-subjects design. A power analysis using G*Power (Faul et al., 2007) based on the effect found in Study 2 (η_p^2 = .032) indicated that we would need at least 396 participants to find the attention effect on surprise with a power of 95%. Accordingly, we set TurkPrime to collect data from 400 participants, 100 per condition. The exclusion criteria were the same as described in Study 2. In total 461 participants started the experiment of which 29 immediately opted out. Another 32 participants met our preregistered fast-rate criteria and were excluded (Greenwald et al., 2003). Six participants failed the attention check embedded in the surprise scale and were also dropped from the final analyses. One person participated twice, and we only included their first set of data, leaving a final sample of 393 participants [9] (48.6% Female; 0.3% non-binary; median age = 34, age range = 18-70 years). Most participants indicated having US-American citizenship (98.7%) and having been born in the USA (94.9%). The majority self-identified as White/Caucasian (74.8%6.6% Black/African American, 4.3% Latino/Hispanic, 5.9% East-Asian, 2.0% South-Asian, 0.5% Middle-Eastern, 5.9% more than one ethnic category or another ethnicity).

Materials

The materials and procedure were based on Study 2, and we manipulated paying attention and predicting IAT scores by omitting individual steps of the prediction procedure. Participants in the attention-prediction condition completed the study exactly as participants in Study 2. They first saw a short explanation of the IAT as a measure of spontaneous affective reactions and were asked to reflect on biases in their reactions while looking at the pictures of Black and White people used on the upcoming IAT. Next, they were asked to rate their reaction by predicting how they would score on a 7-point scale.

In the attention-no-prediction condition, the first part was identical, but the last part was missing. Specifically, the rating scale under the pictures (Step 3) was omitted and replaced by the sentence "Press '>>' when you have thought about what an IAT will show about your spontaneous reactions towards BLACK and WHITE." Hence, these participants were explained that the IAT measures biased affective reactions and then encouraged to pay attention to their biased reactions, but they were never asked to predict the feedback they expected on a scale.

Reversely, participants in the no-attention-prediction condition were only asked to predict how they would score on a test measuring their "implicit attitudes" towards Black and White people, but without explanation of IAT scores reflecting biased affective reactions to pictures of Black and White people, and without encouragement to pay attention to their gut reactions (Steps 1 and 2 were omitted). Participants in the control no-attention-no-prediction condition completed neither of the two steps. Instead, they completed the same filler items as described in Study 2.

Procedure

After providing informed consent and being informed about potential bonus payments, participants were randomly assigned to one of the four conditions and completed the respective tasks. Participants then completed the Black-White IAT (Cronbach's alpha =.75), received feedback based on their performance, and completed the surprise scale (Cronbach's alpha =.96, see Table 1). At the end, all participants filled out demographic information, were given the opportunity to provide feedback on the study, and saw a short debriefing.

Results

Level of Bias

A non-preregistered 2 (attention vs. no-attention) by 2 (prediction scale completion vs. no completion) between-subjects ANOVA on average *D*-scores showed that the manipulations prior to completing the IATs did not significantly influence participants' level of bias. Neither the attention manipulation, $F(1, 389) < 0.01$, $p < .942$, $\eta_p^2 < .001$, nor completing the prediction scale, $F(1, 389)$

[9] With the final sample size of $N = 393$ (randomly assigned to one of four conditions; assuming alpha = 0.05; power = 80%), Study 3 had to reach a minimum effect size of $f = 0.14$ to show a significant effect, which could be reached at a critical F value of 3.87.

= 0.03, $p < .853$, $\eta_p^2 < .001$, nor the interaction, $F(1, 389) = 0.49$, $p < .487$, $\eta_p^2 = .001$ showed significant effects on participants' bias scores.

Surprise at IAT Feedback

A 2 (attention vs. no-attention) by 2 (prediction scale completion vs. no completion) between-subjects ANOVA on average responses on the surprise scale revealed a significant main effect of the attention manipulation, $F(1, 389) = 14.28$, $p < .001$, $\eta_p^2 = .035$ (see Figure 5). The presence of the prediction scale did not have a significant effect on reported surprise, $F(1, 389) < 0.01$, $p = .992$, $\eta_p^2 < .001$. Neither did the data reveal a significant interaction between attention and the presence of the prediction scale, $F(1, 389) = 0.697$, $p = .404$, $\eta_p^2 = .002$.

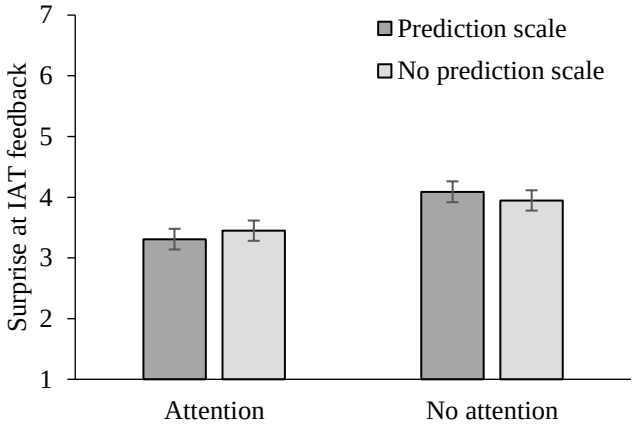
In line with the attention hypothesis, participants who were asked to pay attention to their own affective reactions while looking at pictures used in the IAT reported less surprise at their IAT feedback, independent of whether or not they completed the prediction scale. In contrast, induction to predict (and hence to "admit") one's level of bias on the prediction scale ahead of time had no significant effect on participants' surprise in the absence of encouragement to pay attention to their biases. This last point contradicts the idea that completing the prediction scale shifted participants' perception such that admitting to biases became the more socially desirable response. If that were true, then prediction without attention should have reduced surprise, but this condition showed the most surprise.

Surprise and Strength of Bias Feedback

Replicating findings from Studies 1 and 2, surprise was again uncorrelated with strength of bias feedback, $r(354) = .07$, $p = .174$, nor were there any other significant differences in surprise depending on whether participants received feedback of having "a slight", "a moderate" or "a strong" bias (all pairwise comparisons, ps > .14). However, replicating findings from Study 2 and Howell et al. (2013), participants were again more surprised the more their feedback deviated from their predictions, $r(192) = .65$, $p < .001$.

Figure 5

Study 3: Level of Surprise by Experimental Condition



Note. Level of surprise as a function of paying attention to pictures in IAT score prediction (attention vs. no attention) and using an IAT prediction scale (prediction scale vs. no prediction scale). Error bars depict standard errors of estimated marginal means from a 2 (attention vs. no attention x 2 (prediction scale vs. no prediction scale) ANOVA. N = 393, randomly assigned to condition.

Prediction Accuracy

Within the two prediction conditions, we also looked at the between-subjects correlations between participants' predictions and their IAT scores. In line with Hahn and Goedderz (in prep. a) and as preregistered, people in the present study were more accurate at predicting their level of bias in the attention condition where they saw pictures, $r(97) = .33$, $p = .001$, compared to the no-attention condition without pictures, $r(95) = .17$, $p = .096$. However, in line with observations that differences between correlations require much more power than finding differences between means (Judd et al., 2017), this difference was statistically not significant, $Z = 1.14$ $p = .128$. There was also a non-significant trend for people to predict more bias when they saw pictures ($M = 0.93$, $SD = 1.17$) than when they did not see pictures ($M = 0.60$, $SD = 1.27$), $t(190) = 1.86$, $p = .064$, $d = 0.27$, 95% CI[-0.02; 0.55].

As previously argued (Hahn et al., 2014), one of the problems with between-subjects correlations in discussions around awareness is that such correlations confound awareness with accurate calibration. That is, the degree of a between-subjects correlation does not only depend on whether people are aware of their biases. It also depends on whether they know how biased they are

compared with other people, such that more biased people would have to predict more bias than less biased people. Importantly, the attention hypothesis states that attention makes people recognize their (otherwise preconscious) biases, not that it helps them calibrate them accurately. Following this reasoning, in an additional non-preregistered analysis, we compared whether people more often recognized the direction of their biases (e.g., White over Black, no preference, or Black over White) that showed on the IAT (= 1) or not (= -1)[10], independent of the specific comparative label they chose on the prediction scale. Supporting the attention hypothesis, significantly more people recognized their biases in the attention condition (80.4%) as compared to the no-attention condition (65.3%), X^2 (1, 192) = 5.58, p = .018.

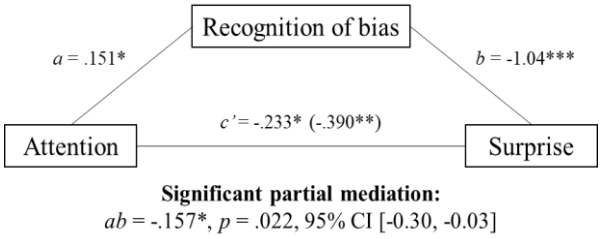
Mediation: Deviation from Expectations or Recognition of Bias?

The analyses so far suggest two possible explanations for why the attention manipulation reduces surprise. First, in line with the feedback wording hypothesis, attention may have lowered participants' surprise only to the degree that their predictions were consistent with the wording used in their IAT feedback. This interpretation is supported by the observation that surprise was a function of deviation of feedback from predictions. However, it stands at odds with the observation that predictions were not significantly more accurate in the attention condition (although this may be a power problem, Judd et al., 2017). A second explanation is in line with the attention hypothesis. It states that attention made people recognize their (formerly unattended) biases, and it was this recognition, independent of the accuracy of their specific predictions, that primarily lowered surprise. To examine these competing explanations, we ran two non-preregistered mediational analyses on participants who predicted their IAT results (Attention-prediction and no-attention-prediction condition; N = 192). Results favored the attention over the feedback wording hypothesis.

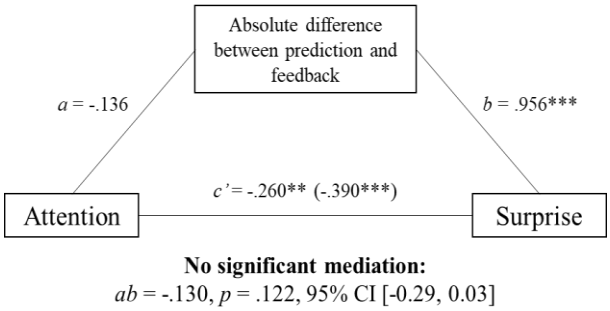
Figure 6, Panel A, shows that an analysis treating recognition of bias as the mediator showed significant mediation.

Figure 6

Study 3: Mediation Analyses with Recognition of Bias (Panel A) or Deviation of Predictions from Feedback (Panel B) Mediating the Effect of Attention on Surprise



A.



B.

Note. Mediation analyses with recognition of bias (Panel A, recognition of bias = 1, no recognition of bias = -1), or the absolute difference between prediction and feedback (Panel B) mediating the effect of attention (attention = 1, no attention = -1) on surprise (scale from 1-7). Bootstrapping analysis indicated significant partial mediation for recognition of bias (Panel A) *ab* = -.157, *p* = .022, *95% CI* [-0.30, -0.03], and no significant mediation for the absolute difference between prediction and feedback (Panel B), *ab* = -.130, *p* = .122, 95% CI [-0.29, 0.03]. Values represent unstandardized path coefficients. The total effect is presented in parentheses.

* indicates significance at the $p < .05$ level, ** at the $p < .01$ level, and *** at the $p < .001$ level.

Participants in the attention condition were significantly more likely to recognize their biases than participants in the no-attention condition (a = .151, p = .018), and participants who recognized their biases were significantly less surprised than participants who did not recognize their biases (b = -1.04, $p < .001$). A bootstrapping analysis for the

[10] Participants were also coded as recognizing their biases (1) when they predicted no preference and showed none, and as not recognizing their biases (-1) when they predicted bias but ended up showing no bias (i.e., |D| < .15).

indirect effect based on 1000 bootstrap samples using R 3.6.1 (R Core Team, 2019) and the *mediation* package (Tingley et al., 2014) revealed a significant partial mediation effect, ab = -.157, p = .022, 95% CI [-0.30, -0.03].

A mediation model treating the absolute difference between predictions and feedback as the mediator did not show evidence for significant mediation (see Figure 6 Panel B). In line with the above reasoning, participants' predictions did not deviate significantly more from their feedback in the attention condition than in the no-attention condition (a = -.135, p = .115), even though deviation of feedback from predictions was related to more surprise (b = .956, p < .001). The bootstrapping analysis using 1000 bootstrap samples revealed no significant mediation, ab = -.130, p = .122, 95% CI [-0.29, 0.03].

Together, these two mediation analyses support the idea that making people pay attention to their spontaneous affective reactions lowers surprise at IAT feedback because it helps them discover their biases. Whether they specifically choose the same labels for their biases as the feedback indicates does not seem to explain lowered surprise in response to predictions.

Discussion

In Study 3, we tested all three competing hypotheses for why IAT score prediction lowers surprise at IAT feedback against each other. Both the attention and the social-desirability hypotheses can explain why prediction lowers surprise, but they predict different mechanisms for this effect. According to the attention hypothesis, predicting IAT scores makes participants pay attention to their biases, which they otherwise rarely consider. In contrast, the social desirability hypothesis states that completing a prediction scale may induce participants to admit to biases they always know. Because revelation of bias is announced during IAT score prediction, and because there is a motivation to be accurate and consistent in one's report (Jussim et al., 1995), prediction may have diminished any value in acting surprised, and shifted participants' perception such that admitting to biases became the socially desirable response.

Results supported the attention hypothesis. People reported less surprise at their IAT feedback when asked to pay attention to their biases prior to IAT completion, even when they never completed the prediction scale. In contrast, merely seeing and completing the prediction scale without an induction to pay attention to one's biases did not make people report less surprise at their IAT feedback. These findings speak against the idea that people are generally aware of their biases but act surprised because such a reaction sounds desirable. Instead, it supports the hypothesis that people discover their biases when they are asked to pay attention to their gut reactions; and this recognition then leads them to report less surprise.

Two non-preregistered mediation analyses additionally showed that recognition of bias was a better explanation for the effect of attention on surprise than lower deviations of predictions from feedback. These analyses provide additional support against the feedback wording hypothesis and in favor of the attention hypothesis. It was recognition of bias, not acceptance of the feedback wording, that explained why attention to biased reactions reduces surprise.

Study 4a

Studies 4a and 4b aimed at examining whether surprise was reduced after predicting IAT results in Studies 2 and 3 because people were encouraged to pay attention to their own reactions (attention hypothesis), or because they read an explanation that the IAT measures spontaneous reactions (Step 1, see Tables 3 and 4, social desirability hypothesis). Specifically, the explanation read as follows: "[…] In addition to the things you say when you are asked about your attitudes, you may have spontaneous reactions toward people at first that you wouldn't always express.[…] For instance, you may have a more positive affective reaction toward a picture of a skinny top model than toward a picture of a regular woman, even though you may not think or say that skinny top models are better people than regular women.[…] [Your implicit attitudes] may be different from the explicit attitudes you would report when you have had time to think about them.".

Reading this information could potentially lead people to be less surprised at their IAT feedback for at least two reasons. First, it may lead them to expect their IAT biases to differ from their explicit attitudes because the explanation says so. Second, the explanation might present IAT results as less of a threat to participants' values and beliefs. Both of those effects may make it less socially

undesirable to admit to bias and report lowered surprise.

To investigate this potential alternative explanation in Study 4a, participants either completed the standard prediction procedure as implemented before; only read the explanation but never predicted their IAT results; or neither read an explanation nor predicted their IAT results, resulting in 3 conditions (explanation and prediction, only explanation, control). Unexpectedly, this study failed to replicate the original prediction effect shown in both Studies 2 and 3. There were no significant differences in reported surprise between conditions, $F(2, 380) = 0.11$, $p = .900$, $\eta_p^2 = .001$.

Because this null-result could be a random false-negative (Lakens & Etz, 2017), and the effect replicated once before, we decided to rerun a slightly altered version of Study 4a, but to include Study 4a in a final mini-meta-analysis (Goh et al., 2016) to investigate whether the main prediction effect still holds when this failed replication is included (see Figure 8). The preregistration for this study can be found at https://osf.io/t5wdn/ and was registered on February 26th, 2019. A more detailed description of the sample and the results can be found in the supplemental materials section. All data, analysis, materials and details on the sample for Study 4a are available on OSF.

Study 4b

As our second attempt to test whether predictions reduced surprise in Studies 2 and 3 due to attention or due to a non-offensive explanation, Study 4b independently manipulated receiving an explanation about implicit attitudes (Step 1) and paying attention to one's reactions by predicting IAT results (Steps 2 and 3) in a 2-by-2 between-subjects design. Hence, Study 4b featured both a condition where people only paid attention to their reactions, but never read any explanation (No-explanation-attention condition), and a condition where they only read an explanation, but never paid attention to their own reactions (Explanation-no-attention condition). See Table 4 for a complete dissociation of the two effects.

Table 4

Components of the prediction procedure included in each condition in Study 4b. The exact wording and all materials can be found on OSF.

		Conditions			
		Explanation		No Explanation	
		Attention	No Attention	Attention	No Attention
		Standard Prediction			Control
	Prediction procedure				
Explanation Manipulation	Step 1: Explanation of IAT as measure of spontaneous affective reactions	Yes	Yes	-	-
Attention Manipulation	Step 2: Pay attention to reaction to pictures used in the IAT Step 3: Completion of prediction scale	Yes	-	Yes	-

Note. The explanation in Step 1 included a non-threatening explanation of implicit attitudes as spontaneous reactions that may be different from explicitly endorsed attitudes. The no-explanation version simply introduced the IAT as a test of "implicit attitudes".

If the attention hypothesis holds true, then this should result in a main effect of the attention manipulation, independent of reading an explanation or not. In contrast, if the social desirability hypothesis is true then this should result in a main effect of the explanation manipulation. In this case, participants who receive the non-threatening introductory text should report less surprise at their IAT results, independent of whether they are asked to pay attention to their biases and predict their IAT scores or not. The preregistration for this study can be found at https://osf.io/h9p6j/ and was registered on April 2nd, 2019.

Method

Participants and Design

Study 4b featured a 2 (explanation vs. no explanation) by 2 (attention vs. no attention) between-subjects design. The attention condition always included predictions of IAT scores based on reactions to pictures. We again aimed at recruiting 400 participants (100 per condition) corresponding to a 95% chance of finding the attention effect from our prior studies. Four-hundred and forty-nine participants started the study, of which 27 instantly opted out. Following the same preregistered exclusion procedure as in Studies 2 and 3, 22 participants were removed due to exceeding speed on the IAT (Greenwald et al., 2003). Another four participants failed to answer the attention check item embedded in our surprise scale, leading to a final sample of 396 participants[11] (54.3% Female; 0.5% non-binary; median age = 35, age range = 18-84 years). Most participants held US-American citizenship (98.2%) and were born in the United States (93.7%). 72.2% of our participants self-identified as White/Caucasian (8.8% Black/African American, 8.1% Latino/Hispanic, 4.3% East-Asian, 1.3% South-Asian, 0.3% Middle-Eastern, 5.1% more than one or none ethnic category).

Materials and Procedure

After being explained the potential bonus payments, participants provided informed consent and were randomly assigned to one of four conditions (see Table 4). Participants in the explanation-attention condition completed the full prediction procedure including all three steps. Participants in the explanation-no-attention condition read the non-threatening intro cited in the introduction to Study 4a, but did not go on to observe their own reactions and predict their scores. Participants in the no-explanation-attention condition only read that they would complete a test that measures their "implicit attitudes" without further information and were then encouraged to observe their reactions towards pictures of Black and White people to predict its results. Participants in the control condition went straight to the IAT without an explanation or predictions.

After finishing the respective tasks, all participants completed a Black-White IAT (Cronbach's alpha = .68), received feedback based on their performance, and were asked to fill out the surprise scale (Cronbach's alpha = .94). Finally, all participants completed a questionnaire on demographic information, were given the chance to provide feedback on the study, and were debriefed.

Results

Level of Bias

A non-preregistered analysis on level of bias as a function of conditions showed that participants' bias scores were not significantly affected by reading the explanation, $F(1, 392) = 1.17$, $p = .279$, $\eta_p^2 = .003$, or paying attention by predicting IAT results, $F(1, 392) = 0.38$, $p < .540$, $\eta_p^2 = .001$. There was no significant interaction between explanation and attention, either, $F(1, 392) < 0.01$, $p = .997$, $\eta_p^2 < .001$.

Surprise at IAT Feedback

We conducted a 2 (explanation vs. no explanation) by 2 (attention vs. no attention) between-subjects ANOVA on average responses on the surprise scale. This analysis supported the attention hypothesis with a significant main effect of attention, $F(1, 392) = 4.05$, $p = .045$, $\eta_p^2 = .010$. Participants who were asked to pay attention to their spontaneous affective reactions toward stimuli by predicting their IAT scores were less surprised at their IAT feedback than participants who did not predict their IAT scores (See Figure 7). Overall, reading the explanation did not

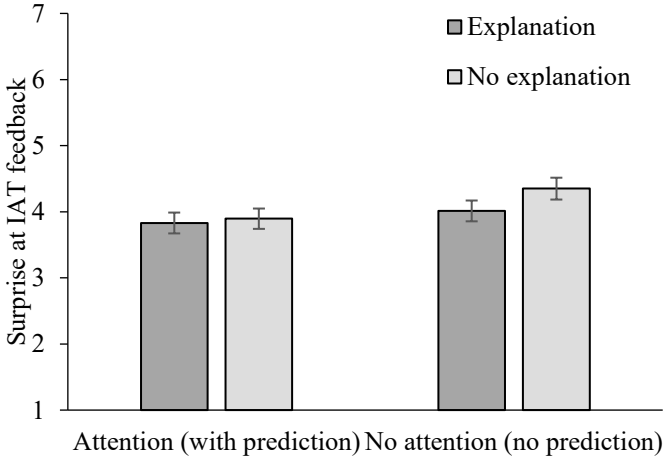
[11] With the final sample size of $N = 396$ (randomly assigned to one of four conditions; assuming alpha = 0.05; power = 80%), Study 4b had to reach a minimum effect size of $f = 0.14$ to find a significant effect, which could be reached at a critical F value of 3.87.

significantly lower surprise, $F(1, 392) = 1.61$, $p = .205$, $\eta_p^2 = .004$, and we did not find a significant interaction either, $F(1, 392) = 0.74$, $p = .391$, $\eta_p^2 = .002$. However, descriptively, participants who only read the explanation were somewhat less surprised than those in the control condition, yet more surprised than those who only paid attention to their reactions but never read any explanation (see Figure 7).

Follow-up simple effects revealed that the attention-only condition differed significantly from the control condition, $F(1, 392) = 3.97$, $p = .047$, $\eta_p^2 = .010$, while the explanation-only condition did not differ significantly from the control condition, $F(1, 392) = 2.23$, $p = .136$, $\eta_p^2 = .006$. At the same time, participants in the explanation-only condition were not significantly more surprised than participants who also paid attention to their reactions, $F(1, 392) = 0.69$, $p = .407$, $\eta_p^2 = .002$. Hence, although the explanation descriptively lowered surprise, the resulting level of surprise fell non-significantly between all other conditions, whereas attention to reactions without explanation did lead to a significant reduction of surprise compared to control.

Figure 7

Study 4b: Level of Surprise by Experimental Condition



Note. Level of surprise as a function of attention to reactions and IAT explanation. Error bars depict standard errors of estimated marginal means from a 2 (attention vs. no attention) x 2 (explanation vs. no explanation) ANOVA. $N = 396$, randomly assigned to condition.

Surprise and Strength of Bias Feedback

Following the same procedure as in Studies 1-3 we looked at the level of surprise as a function of strength of bias feedback, excluding participants who received "little to no" bias feedback. Contrary to our prior findings we found a small but significant correlation between surprise and degree of bias, $r(353) = .12$, $p = .021$. In this study, participants were somewhat more surprised at their feedback the more bias it suggested.

Prediction accuracy

Participants in this study were again able to predict their IAT results, $r(200) = .40$, $p < .001$. Prediction accuracy did not differ as a function of whether participants read an explanation, $r(100) = .39$, $p < .001$, or not, $r(100) = .43$, $p < .001$, $Z = -0.33$ $p = .369$.

Discussion

The purpose of Study 4b was to investigate whether the prediction effect observed in Studies 2 and 3 was due to the fact that participants were encouraged to pay attention to their spontaneous affective reactions (attention hypothesis), or because of the non-threatening and more socially desirable explanation of implicit attitudes as different from explicit attitudes (social desirability hypothesis). To test these two hypotheses against each other, we independently manipulated whether participants read an explanatory text or not, and observed their reactions towards sets of pictures by predicting their IAT results or not. Results were again in line with the attention hypothesis. Participants who were asked to pay attention to their spontaneous affective reactions were less surprised at their IAT feedback independently of whether they read the non-threatening explanation or not. Only reading the explanation did not significantly lower people's surprise compared to the control condition. However, it is noteworthy that, descriptively, participants reported somewhat less surprise at their feedback when provided with a non-threatening explanation; but this explanation alone was neither necessary nor sufficient to lower surprise. And importantly, observing one's reaction without any non-threatening explanation *was* sufficient to reduce surprise.

Unexpectedly, and in contrast to Studies 1-3, participants reported more surprise the harsher the feedback they received. However, this correlation remained low and did not show in any of the other studies, so we treat it as a potential false-positive.

Overall, these findings again support the attention hypothesis and challenge a social

desirability explanation. If people in fact reported less surprise in the previous studies because the non-threatening explanation encouraged them to think that admitting to biases is now desirable, simply reading this explanation should have been sufficient to lower surprise.

Additional Analyses

Meta-Analysis

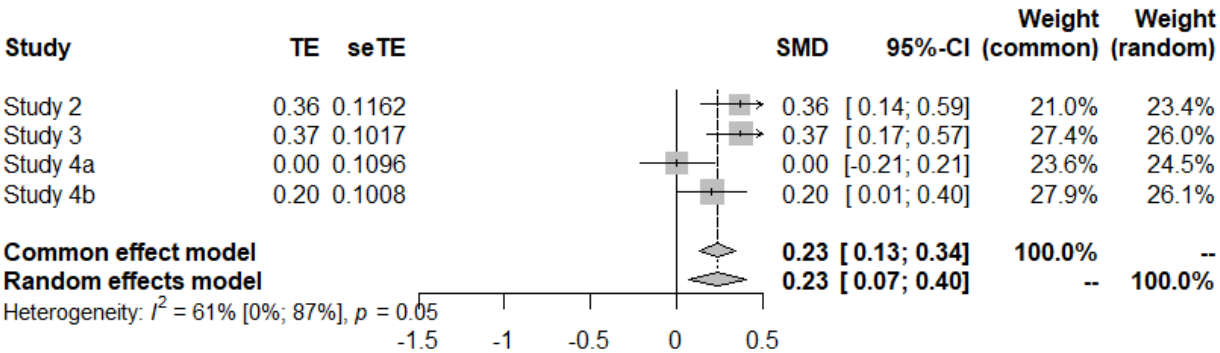
Given the null results of Study 4a, we additionally wanted to investigate whether the overall attention effect would hold across all conducted studies when accounting for the failed replication (Lakens & Etz, 2017). To this end, we meta-analyzed the four studies that included the attention effect (Studies 2-4b, Goh et al., 2016) using R 3.6.1 (R Core Team, 2019) and the *meta* package (Balduzzi et al., 2019). We used a fixed and a random effects model in which the mean effect size (Cohen's *d*) for the main effect of attention in each study was weighted by sample size. Supporting the attention hypothesis, both the fixed and random effects model showed a significant effect for the attention manipulations on surprise including the failed replication (fixed effects model: *M d* = .23, 95% CI [.13, .34]; random effects model: *M d* = .23, 95% CI [.07, .40]; see Figure 8).

Non-White Participants and Participants with Pro-Black Biases

The studies presented in this paper are supposed to present a general effect about intergroup bias, and we hence sampled our participants and preregistered our analyses independent of the majority or minority status of either the participants or the feedback. However, the interested reader may wonder whether White and non-White participants would react similarly in response to receiving pro-Black or pro-White bias feedback, as previous research has shown variations in defensive reactions (e.g., Howell et al., 2015). To investigate this point, we conducted a series of non-preregistered, exploratory analyses across all studies. All results are described in detail in the supplemental materials section. They showed that none of the central results reported in Studies 1-4 interacted with self-reported ethnicity (White vs. non-White) of the participants (all *F*s < 1.2, all *p*s >.29). Additionally, all main results from Studies 1-4 replicated independently on the non-White samples. Furthermore, when combing all relevant data from all studies to yield a more powerful sample, the answers to the three questions from Table 1 showed largely similar patterns on White and non-White participants who showed either pro-White or pro-Black bias separately. Specifically, independently of whether participants were White or non-White, or showed pro-White or pro-Black bias, they always reported more surprise when their feedback indicated any bias (including low levels of bias) than when it indicated no bias (Question 1). In line with findings from Studies 1-3, none of the subsamples showed significant correlations between the reported level of surprise and the strength of the bias feedback (Question 2). Finally, regarding Question 3, whether paying attention to one's biases reduces surprise at IAT feedback, results followed the overall pattern for all subsamples, except for White participants who received pro-Black bias (N = 59). This may suggest that this small sample expected to show pro-White biases after predicting IAT results.

Figure 8

Forest Plot for the Meta-analysis of the Attention Effect over All Conducted Studies



All statistical details on all analyses can be found in the supplemental materials section. Although these results can at best be considered suggestive, given their exploratory nature and the combination of all non-White participants into one sample, they do suggest that many intergroup biases, not just those that indicate undesirable pro-majority bias, but also those that seem less undesirable and may cause less defensive responding (Howell et al., 2015), may be preconscious and hence surprising for many majority and minority group members. More specific and targeted research is needed to investigate and complement these findings.

General Discussion

Awareness that racial biases are widespread across society has been growing (Gallup, 2021). Echoing these developments, social-cognitive research, too, has demonstrated that acknowledgement of implicit bias is possible with simple attention manipulations (Hahn & Gawronski, 2019). At the same time, however, many people seem to react to racial bias feedback on IATs defensively (Howell et al., 2015; Howell et al., 2017; Vitriol & Moskowitz, 2021) and with surprise (Hillard et al., 2013; Schlachter & Rolf, 2017). This can be read as indicating that people do not know that they harbor any racial biases (Gawronski, 2019; Krickel, 2018; Lane et al., 2007), hence supporting the notion that implicit racial bias scores reflect "unconscious" attitudes. The purpose of the current set of studies was to shed light on this apparent contradiction. We proceeded in three steps. First, Study 1 demonstrated the so-far anecdotal observation that participants were in fact more surprised when their feedback indicated bias than when it indicated no bias. Second, Studies 2-4b showed that participants were less surprised at their bias feedback when they predicted their results on an IAT racial bias test prospectively, by looking at the stimuli used on the test and observing their own (biased) reactions. Third and last, Studies 3 and 4 dissociated whether attention, social desirability, or the wording of the feedback were better suited to explain this prediction effect. Results favored the attention hypothesis. Across the four studies, we show that people are surprised at racial bias feedback as long as they are not encouraged to pay attention to their biased reactions. This suggests that the cognitions reflected on implicit measures may often be "preconscious": They are generally accessible, but people rarely pay attention to them (Hahn & Goedderz, 2020).

Alternatively, according to the social desirability hypothesis, people think surprise is the socially desirable response to bias feedback. Completing the prediction slides in the present studies might have reduced participants' reported surprise (but not actual surprise) because the prediction procedure induced them to admit to biases they would otherwise hide. This may have caused them to be prepared for their biases (which they knew all along) to be revealed, and shifted their perception of what a socially desirable response to this revelation would be.

Voicing surprise at anti-Black bias feedback is likely more socially desirable than admitting to harboring biases, and this tendency most certainly contributed to the observed effects. However, our data did not confirm this to be the *main* explanation for the surprise effect. First, surprise was uncorrelated with the social desirability of the feedback. Participants were always more surprised when they received bias feedback as opposed to no-bias feedback, but how much bias the feedback communicated was unrelated to surprise. Second, Study 3 independently manipulated inducing participants to state their biases before the IAT (prediction) and paying attention to spontaneous reactions (attention). Our reasoning was that if the prediction procedure were simply a method to induce participants to "admit" to biases before IAT completion and make this admission seem socially desirable, then prediction without attention should suffice to reduce surprise, whereas attention without prediction should not suffice to reduce surprise. Results did not support this reasoning and instead favored the attention hypothesis. Participants reported less surprise even when they only thought about their affective reactions toward specific stimuli, but never saw or completed a prediction scale. In contrast, merely predicting IAT results on a scale without paying attention to reactions toward pictures did not significantly reduce participants' surprise. Finally, Study 4b showed that the prediction effect was independent of the social desirability of the explanation of the construct of implicit bias. Only paying attention to spontaneous reactions without reading any prior explanation already reduced participants' surprise.

Conversely, when participants only read a non-threatening explanation of the IAT as a measure of spontaneous affective reactions, surprise was not significantly different from either the control condition or the prediction condition. This last finding indicates that non-threatening explanations may help reduce surprise, but not as successfully as active attention to one's reactions.

In sum, our studies did not confirm the idea that participants act surprised at IAT feedback because this answer seems more socially desirable. Instead, it favors the attention explanation: A simple manipulation encouraging people to notice their spontaneous affective reactions made people less surprised at bias feedback. This suggests that people are surprised at IAT feedback because they do not pay attention to their biases chronically – suggesting that these biases are often preconscious (Dehaene et al., 2006).

Yet another explanation for why people might react with surprise at IAT feedback is the feedback wording hypothesis. It says that people disagree with the labels chosen to describe their biases (Gawronski, 2019). Participants in the studies by Hahn et al. (2014) tended to be socially miscalibrated in their results. They disagreed on which biases should be called "mild" or "strong", even as they predicted the patterns of their individual IAT scores accurately. This lack of consensus on what to call a specific reaction might explain why people are surprised at the feedback that they get. Additionally, people are motivated to see themselves as above average on desirable traits and below average on undesirable traits (Alicke et al., 1995). This motivation may lead to surprise at any feedback that suggests that people may have less socially desirable preferences and biases than others.

Empirically investigating this claim led to mixed results. On the one hand, Studies 1-3 did not find any relationship between surprise and the strength of the labels used in the feedback, and Study 4b showed a significant but very small correlation. On the other hand, participants were always more surprised when their feedback included a different qualifier than they had predicted (e.g., "moderate" vs. "strong"). These results do support the notion that surprise is partly a response to the fact that participants have different ideas of what to call their biases than standard IAT feedback communicates. However, it contradicts the notion that the strength or harshness of the feedback is specifically responsible for the surprise reaction. In line with this, a mediation analysis further showed that consistency between predictions and feedback labels did not significantly explain the effect of the attention manipulation on surprise. Instead, it supported the attention hypothesis by showing that the attention manipulation reduced participants' surprise because participants more often recognized their biases compared to the no-attention condition. Lastly, experimentally altering the IAT feedback by omitting all qualifiers and telling all participants with D scores above |.15| that they have "an automatic preference" for one group over the other did not significantly lower surprise compared to standard feedback. In this case, hearing that the IAT suggests "a preference" may have been a bad operationalization for non-threatening feedback; "an automatic preference" might be perceived as more offensive than having a "mild" or "moderate" bias.

Given the remaining ambiguity of these results, we find it important to note that how IAT feedback is communicated may still be an important factor that influences how people react to IATs. For instance Vitriol and Moskowitz (2021) show in their studies that if IAT results are communicated such that participants feel less blamed and perceive more control over their biases, it reduces their defensiveness and increases bias awareness. In the present studies, we only wanted to test the effect of the feedback qualifiers ("slight", "moderate", and "strong") specifically, and found only mixed to negative results. Additional research is needed to further investigate whether communicating feedback in an entirely different way might make people react with less surprise and defensiveness. For instance, instead of using qualifiers that are chosen arbitrarily, it might be helpful to put the feedback in context by telling people where their IAT scores rank in comparison to other people's scores. Additionally, instead of presenting IAT effects as "preferences" for one group over another, they could be framed as "automatic reactions". Such changes in feedback communication might be a more meaningful interpretation of the available data, it could help people understand the meaning of their biases in comparison to others, and it could make them react less defensively to bias feedback.

Yet another explanation for why people react with surprise at IAT results could lie within

problems of the IAT as a measure. That is, task-specific variance of the IAT may have led to distortion of bias scores and thus inaccurate bias feedback. Whether the IAT should be used as a measure of individual bias and thus a basis for individual feedback remains a contentious debate (Kurdi et al., 2020; Schimmack, 2019). However, we believe that task-specific variance cannot explain the attention effect on surprise observed in our studies. First, the fact that participants were able to predict their IAT results suggests that at least in the present studies, the IAT showed a certain degree of validity despite its methodological constraints. Second, if surprise was simply a reaction to "invalid" feedback from the IAT, a prediction manipulation shouldn't have reduced it, since the feedback would have remained just as invalid either way.

In sum, our studies indicate that people are surprised at IAT feedback because they often remain inattentive to their biased reactions towards people. Once they are encouraged to pay attention to these reactions, they discover their biases, and surprise decreases. This suggests that racial biases, such as those reflected on implicit evaluations, are often preconscious. Although they are generally accessible and reportable, people often fail to pay attention to them until they are encouraged to do so.

Limitations

The present studies have several limitations. First, we focused on surprise and awareness of racial biases reflected on IAT scores. While we acknowledge the many criticisms surrounding the IAT and the construct of "implicit bias" (Blanton et al., 2007; Hahn & Gawronski, 2018), however, we believe that our findings may speak to a more general phenomenon. Whether or not the IAT measures them accurately, most humans are likely to hold stereotypic and prejudicial associations with different racial groups that are activated automatically. And independent of whether those automatic biases translate directly to discriminatory behavior (Kurdi et al., 2019), they are likely to show themselves in some responses and reactions. From this perspective, we believe our findings that people tend not to pay attention to their own biases has implications beyond the specific reactions toward IAT feedback we studied here. Although general awareness of racial biases in society may be on the rise (Gallup, 2021), many people may still be resistant to confront their own racial biases, resulting in surprise or even shock at feedback of having shown a bias, even when these biases are easily observable. And this phenomenon might apply to biases beyond those captured by IATs as well. Importantly, as the present studies demonstrate, however, a simple encouragement to pay attention to a biased reaction can change this effect, reduce surprise, and lead to acknowledgement of bias (Hahn & Gawronski, 2019). As we explain below, this insight could prove useful in designing bias intervention and education programs. Future research will have to show how widely our findings that people tend to leave their biases unattended generalizes to other instantiations of racially biased behavior and thoughts.

Second, we only examined racial biases toward Black and White people which may question the generalizability of our findings to other attitudinal domains. We postulate that the specific surprise effect will primarily emerge whenever people's automatic cognitions conflict with their personal standards, such that paying attention to them might be threatening to a person's self-concept. In those cases, encouraging people to pay attention to their reactions before a test should reduce surprise. In contrast, people should be unsurprised at IAT feedback when their explicit evaluations are already based on their spontaneous reactions, such that they match. For instance, Nosek (2007) found high implicit-explicit correspondence in political attitudes, or attitudes towards Coke vs. Pepsi. Because of this correspondence, we would predict little surprise in response to IAT feedback in these domains. Which kinds of attitudes in which domains are subject to such divergences and concerns will likely show variation across cultures, countries, and individuals.

Third, it is important to note that the samples in the reported studies were majority White American participants (71-75%) and the majority of participants received feedback of having a pro-White bias (88- 90%). This limits the generalizability of our findings to other populations. Exploratory analyses across these categories showed that the patterns of results presented in each study and response to our three main questions from Table 1 were similar across White and non-White (American) participants who

showed pro-White or pro-Black bias. While many reactions to intergroup bias feedback – from defensive rejection to emotional responses – will very likely differ as a function of a person's own racial status and the bias in question (e.g., Howell et al., 2015), these results suggest that the specific reaction of surprise may be more general. That is, people may pay little attention to the types of intergroup biases that are reflected on IATs generally. As a result, feedback about them may be surprising independent of whether such feedback is at odds with one's beliefs and values or social desirability, unless one has been encouraged to pay attention. Future research is needed to investigate these questions more directly.

On a more general level, we have reason to believe that our general conclusion – that implicit evaluations often reflect preconscious attitudes to which people may or may not pay attention – will hold across different targets, instruments, and populations. Future research is needed to investigate the generalizability of our findings and conclusions.

Why Are Biases Left Unattended?

One question the present findings pose is how people manage to keep their biases out of awareness (Hahn & Goedderz, 2020). That is, most of the people who participated in these studies have likely met people with different backgrounds in their lives. Why, then, are they discovering new information when asked to observe their reactions? We see several possibilities. The most direct application of the idea of "unattended biases" is that people simply direct away their attention from their biases. Many real-life situations where people may show biases (and thus have a chance to observe them) are ambiguous with respect to the source of the bias, such that identifying a racial bias might need specific motivation. Another, compatible, possibility is that they misattribute their biases to other aspects in the situation. For instance, personal experience with IAT studies suggest that many participants attribute their IAT biases to the order of the blocks in which the IAT is completed. This suggests that they do initially notice their biased reactions, but attribute them to other aspects of the situation than race. This misattribution process might be even more common in real-life situations where there are many more aspects of the situation to which one could attribute one's bias. The current study cannot distinguish between these different interpretations. Importantly, however, regardless of why participants did not pay attention by themselves, once encouraged, participants in this study did notice their biases and reported less surprise at IAT scores. We hope this research contributes to more research on why people are so often blind to biases in their reactions and behavior.

Theoretical Implications for Implicit Bias Research

We believe our studies question the common portrayal of implicit measures as capturing "attitudes people may be unable or unwilling to report" (https://implicit.harvard.edu, 2020). This description summarizes two theoretical perspectives proposed by dual-process models which differ in terms of the role they attribute to consciousness in the dissociation of implicit and explicit measures.. On the one hand, some older conceptualizations have often claimed that implicit evaluations reflect unconscious attitudes people are unable to report (e.g., Greenwald & Banaji, 1995; Lai et al., 2013; Nosek et al., 2002). On the other hand, several dual-process models propose that people might be well-aware of the cognitions reflected on implicit evaluations, but consciously decide not to report those on explicit measures (Fazio, 2007; Fazio & Olson, 2003; Gawronski & Bodenhausen, 2006, 2011).

The present research challenges both perspectives. Assuming that the cognitions reflected on implicit measures are unconscious stands at odds with the present data for at least two reasons. First, people's surprise at IAT feedback was reduced when they paid attention to their biases by predicting IAT scores. If the cognitions reflected in implicit measures were indeed completely unconscious, informing participants that their IAT results might diverge from their explicit attitudes would be the only way to reduce their surprise. However, as Study 4b showed, providing participants with such an explanation was neither sufficient nor necessary to lower people's surprise, while making them pay attention to their biases was both. Second, participants who were asked to pay attention to their biases were overall accurate at predicting their IAT results. Together, these findings speak against the idea that

implicit measures are completely inaccessible to introspection.

However, conceptualizing implicit evaluations as cognitions that are consciously accessible but rejected is also contradicted by the present data. The fact that participants reacted with surprise to IAT feedback in the first place is already hard to reconcile with the idea that people are aware of their biases at all times. Additionally, participants were less surprised after paying attention to their biases. This indicates that they learned new information about their cognitions that they had not considered previously from these predictions. If people were chronically aware of the biases reflected in their implicit evaluations, reduced surprise reactions should be explainable by other factors than paying attention and learning new information, such as the wording of the feedback or social desirability. However, as stated earlier, these explanations cannot fully account for the data.

In sum, the present research speaks against a simple dichotomy of implicit evaluations reflecting either entirely unconscious or entirely conscious cognitions, and thus also against the portrayal of these cognitions as "attitudes people may be unwilling or unable to report" (https://implicit.harvard.edu, 2020). We propose that the concept of "preconsciousness" is better suited to explain the present data and other contradicting findings in the literature on implicit evaluations (Hahn & Goedderz, 2020). A "preconscious" cognition is one that is generally accessible, but to which a person is not paying attention (Dehaene et al., 2006). The present studies suggest that people are surprised at IAT results because, if unattended, the biases captured on implicit evaluations reside outside of conscious awareness. However, those biases are easily accessible when people pay attention to their spontaneous affective reactions, as indicated by reduced surprise reactions and accurate IAT predictions. Hence, although implicit evaluations do not seem to capture "unconscious attitudes" per se, they may well capture cognitions that are preconscious for a lot of people a lot of the time – until they pay attention.

Practical Implications

In addition to these implications for theory, these findings might also have implications for societal debates around implicit bias. There has been a recent debate about the effectiveness of so-called "implicit bias trainings" (e.g., Basu, 2018; Chamorro-Premuzic, 2020; Powell, 2016; Wen, 2020). In light of this trend, we believe refining how implicit evaluations are defined and communicated to the public can ultimately help inform better interventions aimed at reducing biases. While most implicit bias trainings involve many different components and facets that likely need to be approached from different angles, describing implicit biases as "preconscious" and unattended constitutes a much more direct invitation to observe one's own biases than a presentation of bias as "unconscious". And giving people the opportunity to observe their own biases may lead to less defensiveness and hence more openness to the idea of harboring racially biased reactions and behavior.

Additionally, we believe a better and more parsimonious understanding of what implicit biases reflect may benefit societal debates around implicit bias more generally. Our research indicates that implicit biases most likely reflect spontaneous reactions that are often preconscious. Applied to discussions about the meaning of implicit bias for behavior, every person may ask themselves whether they believe their behavior may sometimes be guided by spontaneous, unintentional reactions rather than deliberate attitudes (Hahn & Gawronski, 2018). In contrast to this, discussions on whether implicit evaluations predict behavior have often been presented as discussion around whether behavior is guided entirely by "unconscious forces," or not (e.g., Oswald et al., 2013). This presentation implies that, if the IAT were to predict behavior, we would be helpless victims of mysterious undetectable processes of our mind. Such presentations are not only unlikely to be true, they also appear less likely to lead to solution-oriented discussions than asking people to reflect on their biased impulses. For instance, assumptions about "unconscious forces" guiding our behavior may lead to the conclusion that the only way to correct biased behavior is to reduce implicit biases. However, few interventions aimed at reducing implicit biases have led to long-term changes and may thus not be suited to reduce discriminatory behavior (Lai et al., 2014; Lai et al., 2016). Hence, in addition to adding possible interventions, we hope that our demonstration that the biased cognitions reflected on IAT scores are

often preconscious will contribute to more informed discussions around the possible meaning of implicit biases for society.

Conclusion

Both societal trends and social-psychological research have recently shown that people acknowledge the prevalence of racial biases in society and are able to sense their own racial biases (Gallup, 2021; Hahn et al., 2014; Hahn & Gawronski, 2019). At the same time, researchers report that people who receive feedback of implicit racial bias scores often react defensively and with surprise (Gawronski, 2019; Howell & Ratliff, 2017). To reconcile these seemingly incompatible findings, we proposed the concept of "preconsciousness" (Dehaene et al., 2006). Specifically, we argued that people may often be surprised at their IAT feedback because they rarely pay attention to their racial biases even if in principle, they are accessible. In line with this, the present set of studies first confirmed the so far anecdotal observation that people are surprised at IAT feedback indicating bias. Second, it showed that surprise was reduced when participants were instructed to pay attention to their spontaneous affective reaction before completing a Black-White IAT. Together, these findings show that people often fail to pay attention to their biased reactions. Beyond contributing to our understanding of surprise reactions to racial bias feedback, this also illustrates that going beyond a simple dichotomy of conscious or unconscious attitudes can explain current inconsistencies in research on implicit evaluations, help improve implicit bias interventions, and refine our understanding of implicit social cognition in general.

Many societies around the globe show widespread discrimination and disadvantages for racial minority groups. While there are many reasons for these disparities, one of them might lie in the fact that many well-intentioned people harbor racial biases. While the meaning and origin of these biases will require more research, our studies suggest that simple attention manipulations may be an effective first step towards preventing those biases from being left unattended.

Open Practices

All materials, data sets, and analysis files can be found at https://osf.io/bezkx/.

Preregistrations can be found at the following links:

Study 1: https://osf.io/8uev5/

Study 2: https://osf.io/fh542/

Study 3: https://osf.io/ze3pg/

Study 4a: https://osf.io/t5wdn/

Study 4b: https://osf.io/h9p6j/

References

Alicke, M. D., Klotz, M. L., Breitenbecher, D. L., Yurak, T. J., & al, e. (1995). Personal contact, individuation, and the better-than-average effect. *Journal of Personality and Social Psychology*, *68*(5), 804–825. https://doi.org/10.1037/0022-3514.68.5.804

Balduzzi, S., Rücker, G., & Schwarzer, G. (2019). How to perform a meta-analysis with r: A practical tutorial. *Evidence-Based Mental Health*, *22*(4), 153–160. https://doi.org/10.1136/ebmental-2019-300117

Basu, T. (2018, April 17). *Starbucks will give employees unconscious bias training. That May not help.: when it comes to implicit bias training, the science just doesn't hold up: There isn't much scientific proof that it actually works.* https://www.thedailybeast.com/starbucks-will-give-employees-unconscious-bias-training-that-may-not-help

BBC News. (2017, June 5). *Implicit bias: Is everyone racist?* https://www.bbc.com/news/magazine-40124781

Blanton, H., Jaccard, J., Christie, C., & Gonzales, P. M. (2007). Plausible assumptions, questionable assumptions and post hoc rationalizations: Will the real iat, please stand up? *Journal of Experimental Social Psychology*, *43*(3), 399–409. https://doi.org/10.1016/j.jesp.2006.10.019

Carpenter, T. P., Pogacar, R., Pullig, C., Kouril, M., Aguilar, S., LaBouff, J., Isenberg, N., & Chakroff, A. (2019). Survey-software implicit association tests: A methodological and empirical analysis. *Behavior Research Methods*, *51*(5), 2194–2208. https://doi.org/10.3758/s13428-019-01293-3

Chamorro-Premuzic, T. (2020, January 4). *Implicit bias training doesn't work: instead of changing how employees think, change company policies.* https://www.bloomberg.com/opinion/articles/2020-01-04/implicit-bias-training-isn-t-improving-corporate-diversity

CNN. (2015, November 25). *4 ways you might be displaying hidden bias in everyday life*.

Crowne, D. P., & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology*, *24*(4), 349–354. https://doi.org/10.1037/h0047358

De Houwer, J., Teige-Mocigemba, S., Spruyt, A., & Moors, A. (2009). Implicit measures: A normative analysis and review. *Psychological Bulletin*, *135*(3), 347–368. https://doi.org/10.1037/a0014211

Dehaene, S., Changeux, J.-P., Naccache, L., Sackur, J., & Sergent, C. (2006). Conscious, preconscious, and subliminal processing: A testable taxonomy. *Trends in Cognitive Sciences*, *10*(5), 204–211. https://doi.org/10.1016/j.tics.2006.03.007

Dehaene, S., & Naccache, L. (2001). Towards a cognitive neuroscience of consciousness: Basic evidence and a workspace framework. *Cognition*, *79*(1-2), 1–37.

Dunton, B. C., & Fazio, R. H. (1997). An individual difference measure of motivation to control prejudiced reactions. *Personality and Social Psychology Bulletin*, *23*(3), 316–326. https://doi.org/10.1177/0146167297233009

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.3758/BF03193146

Fazio, R. H. (2007). Attitudes as object-evaluation associations of varying strength. *Social Cognition*, *25*, 603–637. https://doi.org/10.1521/soco.2007.25.5.603

Fazio, R. H., & Olson, M. A. (2003). Implicit measures in social cognition research: Their meaning and use. *Annual Review of Psychology*, *54*(1), 297–327. https://doi.org/10.1146/annurev.psych.54.101601.145225

Gallup. (2021). *Race relations.* Gallup. https://news.gallup.com/poll/1687/race-relations.aspx

Gawronski, B. (2019). Six lessons for a cogent science of implicit bias and its criticism. *Perspectives on Psychological Science : A Journal of the Association for Psychological Science*, *14*(4), 574–595. https://doi.org/10.1177/1745691619826015

Gawronski, B., & Bodenhausen, G. V. (2006). Associative and propositional processes in evaluation: An integrative review of implicit and explicit attitude change. *Psychological Bulletin*, *132*(5), 692–731. https://doi.org/10.1037/0033-2909.132.5.692

Gawronski, B., & Bodenhausen, G. V. (2011). The associative-propositional evaluation model: Theory, evidence, and open questions. *Advances in Experimental Social Psychology*, *44*, 59–127. https://doi.org/10.1016/B978-0-12-385522-0.00002-0

Gawronski, B., & De Houwer, J. (2014). Implicit measures in social and personality psychology. In H. T. Reis & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (2nd ed., pp. 283–310). Cambridge University Press.

Gawronski, B., Hofmann, W., & Wilbur, C. J. (2006). Are "implicit" attitudes unconscious? *Consciousness and Cognition*, *15*(3), 485–499. https://doi.org/10.1016/j.concog.2005.11.007

Goh, J. X., Hall, J. A., & Rosenthal, R. (2016). Mini meta-analysis of your own studies: Some arguments on why and a primer on how. *Social and Personality Psychology Compass*, *10*(10), 535–549. https://doi.org/10.1111/spc3.12267

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes.

Psychological Review, *102*(1), 4–27. https://doi.org/10.1037/0033-295X.102.1.4

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology*, *74*, 1464–1480. https://doi.org/10.1037/0022-3514.74.6.1464

Greenwald, A. G., Nosek, B. A., & Banaji, M. R. (2003). Understanding and using the implicit association test: I. An improved scoring algorithm. *Journal of Personality and Social Psychology*, *85*(2), 197–216. https://doi.org/10.1037/0022-3514.85.2.197

The Guardian. (2018, December 2). *Unconscious bias: What is it and can it be eliminated?* https://www.theguardian.com/uk-news/2018/dec/02/unconscious-bias-what-is-it-and-can-it-be-eliminated

The Guardian. (2021, April 25). *What unconscious bias training gets wrong… and how to fix it*. https://www.theguardian.com/science/2021/apr/25/what-unconscious-bias-training-gets-wrong-and-how-to-fix-it

Hahn, A., & Gawronski, B. (2014). Do implicit evaluations reflect unconscious attitudes? *Behavioral and Brain Sciences*, *37*(1), 28–29. https://doi.org/10.1017/S0140525X13000721

Hahn, A., & Gawronski, B. (2018). Implicit social cognition. *Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience*, *4*, 1–33. https://onlinelibrary.wiley.com/doi/full/10.1002/9781119170174.epcn412

Hahn, A., & Gawronski, B. (2019). Facing one's implicit biases: From awareness to acknowledgment. *Journal of Personality and Social Psychology*, *116*(5), 769–794. https://doi.org/10.1037/pspi0000155

Hahn, A., & Goedderz, A. (in prep. a). *Beyond dishonesty and unawareness. Accuracy of IAT score predictions depends more on the concreteness of the question than on knowledge of measurement* [Manuscript in preparation]. University of Cologne.

Hahn, A., & Goedderz, A. (2020). Trait-unconsciousness, state-unconsciousness, preconsciousness, and social miscalibration in the context of implicit evaluation. *Social Cognition*, *38*(Supplement), s115-s134. https://doi.org/10.1521/soco.2020.38.supp.s115

Hahn, A., Judd, C. M., Hirsh, H. K., & Blair, I. V. (2014). Awareness of implicit attitudes. *Journal of Experimental Psychology: General*, *143*(3), 1369–1392. https://doi.org/10.1037/a0035028

Haider, A. H., Schneider, E. B., Sriram, N., Dossick, D. S., Scott, V. K., Swoboda, S. M., Losonczy, L., Haut, E. R., Efron, D. T., Pronovost, P. J., Freischlag, J. A., Lipsett, P. A., Cornwell, E. E., MacKenzie, E. J., & Cooper, L. A. (2014). Unconscious race and class bias: Its association with decision making by trauma and acute care surgeons. *The Journal of Trauma and Acute Care Surgery*, *77*(3), 409–416. https://doi.org/10.1097/TA.0000000000000392

Hillard, A. L., Ryan, C. S., & Gervais, S. J. (2013). Reactions to the implicit association test as an educational tool: A mixed methods study. *Social Psychology of Education*, *16*(3), 495–516. https://doi.org/10.1007/s11218-013-9219-5

Hofmann, W., Gawronski, B., Gschwendner, T., Le, H., & Schmitt, M. (2005). A meta-analysis on the correlation between the implicit association test and explicit self-report measures. *Personality and Social Psychology Bulletin*, *31*(10), 1369–1385. https://doi.org/10.1177/0146167205275613

Hofmann, W., Gschwendner, T., Nosek, B. A., & Schmitt, M. (2005). What moderates implicit—explicit consistency? *European Review of Social Psychology*, *16*(1), 335–390. https://doi.org/10.1080/10463280500443228

Hofmann, W., Gschwendner, T., & Schmitt, M. (2009). The road to the unconscious self not taken: Discrepancies between self- and observer-inferences about implicit dispositions from nonverbal behavioural cues. *European Journal of Personality*, *23*(4), 343–366. https://doi.org/10.1002/per.722

Hofmann, W., & Wilson, T. D. (2010). Consciousness, introspection, and the adaptive unconscious. In B. Gawronski & B. K. Payne (Eds.), *Handbook of implicit social cognition: Measurement, theory, and applications* (pp. 197–215). Guildford Press.

Howell, J. L., Collisson, B., Crysel, L., Garrido, C. O., Newell, S. M., Cottrell, C. A., Smith, C. T., & Shepperd, J. A. (2013). Managing the threat of impending implicit attitude feedback. *Social Psychological and Personality Science*, *4*(6), 714–720. https://doi.org/10.1177/1948550613479803

Howell, J. L., Gaither, S. E., & Ratliff, K. A. (2015). Caught in the middle: Defensive responses to iat feedback among whites, blacks, and biracial black/whites. *Social Psychological and Personality Science*, *6*(4), 373–381. https://doi.org/10.1177/1948550614561127

Howell, J. L., & Ratliff, K. A. (2017). Not your average bigot: The better-than-average effect and defensive responding to implicit association test feedback. *British Journal of Social Psychology*, *56*(1), 125–145. https://doi.org/10.1111/bjso.12168

Howell, J. L., Redford, L., Pogge, G., & Ratliff, K. A. (2017). Defensive responding to iat feedback. *Social Cognition*, *35*(5), 520–562. https://doi.org/10.1521/soco.2017.35.5.520

https://implicit.harvard.edu. (2020).

Judd, C. M., McClelland, G. H., & Ryan, C. S. (2017). *Data analysis: A model comparison approach to regression, ANOVA, and beyond* (Third Edition). Routledge Taylor & Francis Group.

Jussim, L., Yen, H., & Aiello, J. R. (1995). Self-consistency, self-enhancement, and accuracy in reactions to feedback.

Journal of Experimental Social Psychology, *31*(4), 322–356. https://doi.org/10.1006/jesp.1995.1015

Krickel, B. (2018). Are the states underlying implicit biases unconscious? – a neo-freudian answer. *Philosophical Psychology*, *31*(7), 1007–1026. https://doi.org/10.1080/09515089.2018.1470323

Kurdi, B., Ratliff, K. A., & Cunningham, W. A. (2020). Can the implicit association test serve as a valid measure of automatic cognition? A response to schimmack (2020). *Perspectives on Psychological Science*, 1-13. https://doi.org/10.1177/1745691620904080

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. *The American Psychologist*, *74*(5), 569–586. https://doi.org/10.1037/amp0000364

Lai, C. K., Hoffman, K. M., & Nosek, B. A. (2013). Reducing implicit prejudice. *Social and Personality Psychology Compass*, *7*(5), 315–330. https://doi.org/10.1111/spc3.12023

Lai, C. K., Marini, M., Lehr, S. A., Cerruti, C., Shin, J.-E. L., Joy-Gaba, J. A., Ho, A. K., Teachman, B. A., Wojcik, S. P., Koleva, S. P., Frazier, R. S., Heiphetz, L., Chen, E. E., Turner, R. N., Haidt, J., Kesebir, S., Hawkins, C. B., Schaefer, H. S., Rubichi, S., . . . Nosek, B. A. (2014). Reducing implicit racial preferences: I. A comparative investigation of 17 interventions. *Journal of Experimental Psychology. General*, *143*(4), 1765–1785. https://doi.org/10.1037/a0036260

Lai, C. K., Skinner, A. L., Cooley, E., Murrar, S., Brauer, M., Devos, T., Calanchini, J., Xiao, Y. J., Pedram, C., Marshburn, C. K., Simon, S., Blanchar, J. C., Joy-Gaba, J. A., Conway, J., Redford, L., Klein, R. A., Roussos, G., Schellhaas, F. M. H., Burns, M., . . . Nosek, B. A. (2016). Reducing implicit racial preferences: Ii. Intervention effectiveness across time. *Journal of Experimental Psychology. General*, *145*(8), 1001–1016. https://doi.org/10.1037/xge0000179

Lakens, D., & Etz, A. J. (2017). Too true to be bad: When sets of studies with significant and nonsignificant findings are probably true. *Social Psychological and Personality Science*, *8*(8), 875–881. https://doi.org/10.1177/1948550617693058

Lane, K. A., Banaji, M. R., Nosek, B. A., & Greenwald, A. G. (2007). Understanding the implicit association test: Iv. *Implicit Measures of Attitudes*, 59–102.

Lindsay, D. S., Simons, D. J., & Lilienfeld, S. O. (2016, November 30). *Research preregistration 101*. https://www.psychologicalscience.org/observer/research-preregistration-101

Minear, M., & Park, D. C. (2004). A lifespan database of adult facial stimuli. *Behavior Research Methods, Instruments, & Computers*, *36*(4), 630–633. https://doi.org/10.3758/BF03206543

Nosek, B. A. (2005). Moderators of the relationship between implicit and explicit evaluation. *Journal of Experimental Psychology: General*, *134*(4), 565–584. https://doi.org/10.1037/0096-3445.134.4.565

Nosek, B. A. (2007). Implicit–explicit relations. *Current Directions in Psychological Science*, *16*(2), 65–69. https://doi.org/10.1111/j.1467-8721.2007.00477.x

Nosek, B. A., Banaji, M., & Greenwald, A. G. (2002). Harvesting implicit group attitudes and beliefs from a demonstration web site. *Group Dynamics: Theory, Research, and Practice*, *6*(1), 101–115. https://doi.org/10.1037//1089-2699.6.1.101

Nosek, B. A., & Hansen, J. J. (2008). The associations in our heads belong to US: Searching for attitudes and knowledge in implicit evaluation. *Cognition & Emotion*, *22*(4), 553–594. https://doi.org/10.1080/02699930701438186

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*(2), 171–192. https://doi.org/10.1037/a0032734

Perry, S. P., Murphy, M. C., & Dovidio, J. F. (2015). Modern prejudice: Subtle, but unconscious? The role of Bias Awareness in Whites' perceptions of personal and others' biases. *Journal of Experimental Social Psychology*, *61*, 64–78. https://doi.org/10.1016/j.jesp.2015.06.007

Powell, J. A. (2016, September 27). *Implicit bias in the presidential debate [blog post]*. UC Berkeley. Berkeley Blog. https://blogs.berkeley.edu/2016/09/27/implicit-bias-in-the-presidential-debate/

Quillian, L. (2008). Does unconscious racism exist? *Social Psychology Quarterly*, *71*(1), 6–11. https://doi.org/10.1177/019027250807100103

Ranganath, K. A., Smith, C. T., & Nosek, B. A. (2008). Distinguishing automatic and controlled components of attitudes from direct and indirect measurement methods. *Journal of Experimental Social Psychology*, *44*(2), 386–396. https://doi.org/10.1016/j.jesp.2006.12.008

Redford, L. (2018). *Mapping county-level geographical variation in implicit racial attitudes* [November 4, 2018]. Project Implicit. https://www.implicit.harvard.edu/implicit/blog.html

Schimmack, U. (2019). The implicit association test: A method in search of a construct. *Perspectives on Psychological Science.* Advance online publication. https://doi.org/10.1177/1745691619863798

Schlachter, S., & Rolf, S. (2017). Using the iat: How do individuals respond to their results? *International Journal of Social Research Methodology*, *20*(1), 77–92. https://doi.org/10.1080/13645579.2015.1117799

Scientific American. (2020, August 28). *The problem with implicit bias training*. https://www.scientificamerican.com/article/the-problem-with-implicit-bias-training/

Sedikides, C., Gaertner, L., & Toguchi, Y. (2003). Pancultural self-enhancement. *Journal of Personality and Social Psychology*, *84*(1), 60–79. https://doi.org/10.1037/0022-3514.84.1.60

Smith, C. T., & Nosek, B. A. (2011). Affective focus increases the concordance between implicit and explicit attitudes. *Social Psychology*, *42*(4), 300–313. https://doi.org/10.1027/1864-9335/a000072

Stiensmeier-pelster, J., Martini, A., & Reisenzein, R. (1995). The role of surprise in the attribution process. *Cognition & Emotion*, *9*(1), 5–31. https://doi.org/10.1080/02699939508408963

Tingley, D., Yamamoto, T., Hirose, K., Keele, L., & Imai, K. (2014). Mediation: R package for causal mediation analysis. *Journal of Statistical Software*, *59*(5). http://www.jstatsoft.org/v59/i05/

Vitriol, J. A., & Moskowitz, G. B. (2021). Reducing defensive responding to implicit bias feedback: On the role of perceived moral threat and efficacy to change. Journal of Experimental Social Psychology, 96(8), 104165. https://doi.org/10.1016/j.jesp.2021.104165

Wen, T. (2020, August 28). *Is it possible to rid police officers of bias?* https://www.bbc.com/future/article/20200827-is-it-possible-to-rid-police-officers-of-bias