American Psychological Association

Journal of Personality and Social Psychology:
Interpersonal Relations and Group Processes

0022-3514/18/$12.00

http://dx.doi.org/10.1037/pspi0000155

Facing One's Implicit Biases: From Awareness to Acknowledgment

Adam Hahn
University of Cologne

Bertram Gawronski
University of Texas at Austin

Expanding on conflicting theoretical conceptualizations of implicit bias, 6 studies tested the effectiveness of different procedures to increase acknowledgment of harboring biases against minorities. Participants who predicted their responses toward pictures of various minority groups on future implicit association tests (IATs) showed increased alignment between implicit and explicit preferences (Studies 1–3), greater levels of explicit bias (Studies 1–3), and increased self-reported acknowledgment of being racially biased (Studies 4–6). In all studies, effects of IAT score prediction were significant even when participants did not actually complete IATs. Effects of predicting IAT scores were moderated by nonprejudicial goals, in that IAT score prediction increased acknowledgment of bias for participants with strong nonprejudicial goals, but not for participants with weak nonprejudicial goals (Study 4). Mere completion of IATs and feedback on IAT performance had inconsistent effects across studies and criterion measures. Instructions to attend to one's spontaneous affective reactions toward minority group members increased acknowledgment of bias to the same extent as IAT score prediction (Study 6). The findings are consistent with conceptualizations suggesting that (a) implicit evaluations are consciously experienced as spontaneous affective reactions and (b) directing people's attention to their spontaneous affective reactions can increase acknowledgment of bias. Implications for theoretical conceptualizations of implicit bias and interventions that aim to reduce discrimination via increased acknowledgment of bias are discussed.

Keywords: awareness, implicit association test, intervention, implicit bias, prejudice reduction

Supplemental materials: http://dx.doi.org/10.1037/pspi0000155.supp

It is often assumed that, in order to counteract discrimination, people must acknowledge that they harbor intergroup biases such as those reflected in implicit evaluations. Performance-based measures such as the implicit association test (IAT; Greenwald, McGhee, & Schwartz, 1998), the evaluative priming task (EPT; Fazio, Jackson, Dunton, & Williams, 1995), and the affect misattribution procedure (AMP; Payne, Cheng, Govorun, & Stewart, 2005) have revealed that people may harbor implicit biases even when they do not show explicit biases on traditional self-report measures (e.g., Nosek, Banaji, & Greenwald, 2002). Many researchers and policymakers have responded to these findings with initiatives to educate the public about implicit biases, assuming that acknowledgment of these biases is an important step in counteracting discrimination. In line with this idea, acknowledgment of bias plays a central role in theories of prejudice reduction (e.g., Monteith & Mark, 2005) and may have contributed to efforts to make implicit bias tests widely accessible through online tools (e.g., http://implicit.harvard.edu). The popular media are similarly replete with anecdotal reports of the positive effects of learning about one's implicit biases (e.g., Dateline NBC, 2007; This American Life, 2015) and calls to make implicit bias testing mandatory for certain professions (e.g., Reuters, 2016). In fact, then-presidential candidate Hillary Clinton announced she would dedicate funds to implicit bias training for police officers and other professions if she won the 2016 U.S. Presidential Election (Hillaryclinton.com, 2016). However, despite the central role commonly attributed to acknowledgment of bias, the factors that lead to such acknowledgment are still not well understood. Moreover, different theoretical conceptions of implicit bias suggest different interventions to increase acknowledgment of bias, but scientific consensus about the validity of these conceptions is still lacking.

Expanding on conflicting conceptions of implicit bias, the current research investigated the effectiveness of different procedures to increase acknowledgment of personal bias. Toward this end, we tested the extent to which acknowledgment of bias is increased by predicting one's IAT scores (Hahn, Judd, Hirsh, & Blair, 2014)

Adam Hahn, Social Cognition Center Cologne, Department of Psychology, University of Cologne; Bertram Gawronski, Department of Psychology, University of Texas at Austin.

We thank our many research assistants for help with programming and running these studies. Special thanks go to Timo Luoma and Nicole Hartwich (program and administration, Study 2), Clara Nicolina Bersch and Timo Luoma (program and administration, Study 3), and Alexandra Goeddarz (programs, Studies 4 and 5), as well as to the many research assistants taking on the role of experimenters. Study 1 was supported by a Canada Research Chair grant awarded to Bertram Gawronski. Study 3 was funded with a grant by the Center of Social and Economic Behavior (C-SEB) at the University of Cologne awarded to Adam Hahn. Study 5 was funded by grant by Deutsche Forschungsgesellschaft (DFG, German Research Society) awarded to Adam Hahn.

Correspondence concerning this article should be addressed to Adam Hahn, Social Cognition Center Cologne, Department of Psychology, University of Cologne, Richard-Strauss-Straße 2, 50931 Köln, Germany. E-mail: Adam.Hahn@uni-koeln.de

and compared the obtained effects with those resulting from actual IAT completion and feedback about one's IAT performance.

Theories of Implicit Evaluation and Acknowledgment of Bias

Implicit evaluations are often assumed to reflect attitudes and beliefs people are either unable or unwilling to report (e.g., https://implicit.harvard.edu/implicit/education.html). Although the two conceptions are frequently mentioned within the same sentence, they suggest different ways to increase acknowledgment of bias.

The idea that people are unable to report their implicit biases is captured by the notion that implicit evaluations reflect unconscious attitudes to which people have no introspective access (for critical reviews, see Gawronski, Hofmann, & Wilbur, 2006; Hahn & Gawronski, 2014; Hahn et al., 2014). Based on Greenwald and Banaji's (1995) definition of implicit attitudes as "introspectively unidentified (or inaccurately identified) traces of past experience that mediate favorable or unfavorable feeling, thought, or action toward social objects" (p. 8), some researchers argue that people have no introspective access to their implicit biases, which makes them unable to report these biases on traditional self-report measures. From this perspective, one potential way to increase acknowledgment of bias is to inform people about their implicit biases by providing individual feedback on their measurement scores. For example, after completing an IAT on the Project Implicit web page (http://implicit.harvard.edu), participants typically receive feedback on their personal level of bias, which is derived from numeric cut-offs that translate a participant's IAT score into verbal feedback (e.g., a *D*-score of 0.2 on the race IAT reflecting *a slight automatic preference for White over Black*). Applied to the current question, such feedback may increase acknowledgment of bias by informing participants about their unconscious attitudes, which they are unable to know without feedback on their personal measurement scores.

The idea that people are unwilling to report their implicit biases is captured by the notion that dissociations between implicit and explicit evaluations reflect self-presentational distortions on self-report measures (for a critical review, see Gawronski, LeBel, & Peters, 2007). Because responses on traditional self-report measures are much easier to control than responses on performance-based measures, honest reporting of one's biases on self-report measures can be undermined by social desirability and other self-presentational concerns (Dunton, & Fazio, 1997). From this perspective, one potential way to increase acknowledgment of bias is to create contexts that encourage participants to openly admit to their biases. One possibility may be to tell participants that their personal biases will be identified with a performance-based measure that cannot be controlled (e.g., Nier, 2005). The underlying idea is that announcement of a test should make intentional misreporting futile. Applied to the current question, knowledge that one's biases will be uncovered by a psychological test may increase people's willingness to admit to their personal biases, and this increase may occur without individual feedback on one's measurement scores from a bias test.

A third perspective assumes that people have no introspective access to the attitudes underlying implicit biases, but people can become aware of these attitudes by observing their behavioral effects (Hofmann, Gschwendner, & Schmitt, 2009; Hofmann & Wilson, 2010). Although performance-based measures differ in the extent to which participants become aware of systematic differences in their behavioral responses (Petty, Fazio, & Briñol, 2009), participants completing the IAT often notice the difference in their reaction times (RTs) and errors in the prejudice-congruent and prejudice-incongruent blocks (Monteith, Voils, & Ashburn-Nardo, 2001). In line with the notion of self-perception (Bem, 1972) this perspective suggests that mere completion of an IAT even without feedback about one's measurement scores may increase acknowledgment of bias to the extent that participants notice the behavioral effects of their attitudes in the task.

A fourth conception suggests that implicit evaluations are subjectively experienced as spontaneous affective reactions, and dissociations between implicit and explicit evaluations arise from differences in the extent to which people rely on their spontaneous affective reactions in making an evaluative judgment (Fazio, 2007; Gawronski & Bodenhausen, 2006, 2011). To the extent that people accept their spontaneous affective reactions as a basis for evaluative judgments, implicit and explicit evaluations are assumed to align. In contrast, if people reject their spontaneous affective reactions, implicit and explicit evaluations should dissociate. From this perspective, directing people's attention to their spontaneous affective reactions toward minority members may increase acknowledgment of bias by counteracting the dismissal of these reactions as a basis for an explicit judgment. Different from the previous three conceptions, this view suggests that acknowledgment of bias could be increased by directing people's attention to their spontaneous affective reactions without requiring them to complete an IAT, without feedback on their measurement scores, and without anticipation of actual measurement.

In the current research, we tested the effectiveness of these theoretically derived procedures in increasing acknowledgment of bias, focusing particularly on the prediction of one's IAT scores as a procedure to enhance attention to one's spontaneous affective reactions (see Hahn et al., 2014). Toward this end, participants predicted their IAT scores or not, completed IATs or not, and received feedback on their IAT performance or not. In addition, we compared effects of IAT score prediction with the effects of merely attending to one's spontaneous affective reactions toward minority groups. The overarching goal was to shed light on which of these strategies would increase acknowledgment of bias as reflected in (a) greater alignment between implicit and explicit preferences, (b) greater levels of explicit bias, and (c) greater self-reported acknowledgment of being biased.

IAT Completion and Acknowledgment of Bias

Previous research on whether IAT feedback leads people to acknowledge their biases has revealed mixed evidence. Analyzing over 100,000 responses to IAT feedback on the IAT website, Howell and colleagues (Howell, Gaither, & Ratliff, 2015; Howell & Ratliff, 2017) concluded that participants tend to react defensively when their feedback deviates from their performance expectations. Participants whose IAT feedback suggested more bias than they had ascribed to themselves were more likely to question the validity of the IAT than participants whose IAT feedback confirmed their contentions. These findings suggest that participants who could learn the most from their feedback are the ones who are most likely to reject it; only participants who receive

confirmatory feedback seem to accept it. Echoing this concern, other research has shown that a substantial number of people attribute their performance on the IAT to factors other than personal bias (e.g., Casad, Flores, & Didway, 2013; Monteith et al., 2001; Uhlmann & Nosek, 2012). A potential explanation for these attributions is that completing racial IATs can be a threatening experience for White participants, particularly when these IATs are described as diagnostic measures of racism (Frantz, Cuddy, Burnett, Ray, & Hart, 2004).

Deviating from the approach of the IAT website where participants receive feedback on completed IATs, Monteith et al. (2001) tested whether participants notice their performance differences on different blocks of the IAT. In their study, 64% did notice such differences, but only 37% of these participants attributed them to race-related factors as opposed to race-unrelated factors (e.g., block order, color associations). Participants who noticed their performance differences and attributed them to racial bias tended to experience negative affect, which the authors viewed as an important step in enhancing the motivation to control prejudiced reactions. These findings demonstrate that recognizing bias from IAT completion is possible and may have downstream effects on self-perception. However, these effects seem to be limited to a relatively small proportion of participants.

Research adopting a "bogus pipeline" approach (Nier, 2005) found higher correlations between IAT scores and the Modern Racism Scale (MRS, McConahay, 1986) when participants were told that (a) the IAT is a valid measure of racial attitudes akin to a "lie detector" and (b) they should complete the MRS as if the computer on which they had completed the IAT predicted their responses. Correlations between the IAT and the MRS were relatively low when participants completed the IAT without this information or with information questioning the validity of the IAT. These results are consistent with the idea that anticipating identification of one's biases with a psychological test may increase people's willingness to admit their personal biases. However, one may question whether instructions to complete a self-report measure "like the computer would" is equivalent to acknowledgment of harboring racial biases.

Research on classroom use of the IAT (Casad et al., 2013; Hillard, Ryan, & Gervais, 2013), as well as anecdotal evidence from bias awareness trainings (e.g., Dateline NBC, 2007; This American Life, 2015) generally report positive effects of completing IAT measures of implicit bias and discussing their outcomes. However, all of these studies lacked suitable control conditions or explicit bias measures prior to completing the IAT. Hence, it remains unclear whether high correlations between IAT scores and post-IAT measures of explicit bias (which may reflect "acknowledgment of bias") are the result of completing IATs (Hillard et al., 2013). Similarly, findings that participants experience more positive than negative affect after an IAT class exercise (Hillard et al., 2013) are difficult to interpret when there is no control condition to assess participants' baseline affect. After all, most people feel more positive than negative affect most of the time (Watson, Clark, & Tellegen, 1988). Even if one were to accept the conclusion that these classroom and awareness exercises lead to increased acknowledgment of bias, it would remain unclear if the obtained effects are caused by the completion of IATs or by other aspects of the exercises, such as reflection about one's personal biases.

IAT Score Prediction and Acknowledgment of Bias

Counter to the widespread assumption that implicit biases reflect unconscious attitudes, Hahn, Judd, Hirsh, and Blair (2014) found that people can predict their patterns of scores on prejudice IATs with a high degree of accuracy. This result differs from findings by Howell and colleagues (Howell et al., 2015; Howell & Ratliff, 2017) showing that people chose less negative labels to describe their biases than the feedback they received on the IAT website and responded defensively to negative feedback. Although people may not believe that their bias for one group over another is "very strong" (and react defensively if a computer program tells them so), they may still be able to identify with a high degree of accuracy whether they show more bias against some groups than others (see Hahn et al., 2014, for a more detailed discussion on this difference). Interestingly, participants who accurately predicted the patterns of their IAT scores and then completed those IATs later showed (a) greater explicit preferences for Whites over minorities and (b) greater alignment between explicit and implicit preferences. Together, these findings suggest that predicting one's IAT scores may be more effective in increasing acknowledgment of bias than actual IAT completion and IAT feedback. However, at least two points make this conclusion premature. First, Hahn et al.'s (2014) findings do not address the question of whether the obtained changes in explicit bias are caused by the prediction of IAT scores, actual IAT completion, or a combination of the two. Second, it remains unclear whether greater explicit preference for Whites over minorities and greater alignment of explicit and implicit preferences reflects genuine acknowledgment of bias. The purpose of the current research is to shed light on these questions. Below we elaborate on each question before we present the two approaches with which we addressed them.

IAT Completion Versus IAT Score Prediction

Hahn et al.'s (2014) finding that people can predict their IAT scores with a high degree of accuracy poses a challenge to the idea that implicit biases reflect unconscious attitudes that people are unable to report. Nevertheless, the obtained increase in explicit biases and the increased alignment between implicit and explicit preferences are compatible with three of the ideas mentioned above on how to increase acknowledgment of bias. First, because participants were told that they would have to complete the IATs for which they were asked to make predictions, the obtained effects are consistent with the hypothesis that acknowledgment of bias can be increased by creating a context that encourages participants to openly admit to their biases. According to this view, anticipating that one's biases will be uncovered by a psychological test may increase people's willingness to admit to their personal biases because it would be futile to continue concealing them, and this increase may occur without actual completion of an implicit bias test or individual feedback on one's measurement scores. Second, because participants actually completed IATs, the obtained effects are consistent with the hypothesis that people may become aware of their unconscious attitudes by observing behavioral effects of these attitudes while completing IATs. Finally, because participants were asked to predict their IAT scores, the obtained effects are consistent with the hypothesis that directing people's attention to their spontaneous affective reactions may increase acknowledgment of bias by counteracting the dismissal of

these reactions. According to this view, the prediction task may direct people's attention to their spontaneous affective reactions, which should increase acknowledgment of bias without actual completion of an IAT, and without anticipation of actual measurement. Based on these considerations, the obtained increase in explicit biases and the increased alignment between implicit and explicit preferences could be the result of any of the three mechanisms (or any of their combinations).

Acknowledgment of Bias?

Another important question is whether the obtained increase in explicit biases and greater alignment between implicit and explicit preferences reflect increased acknowledgment of bias. In Hahn et al.'s (2014) studies, participants reported greater negativity toward Latinos, Asians, and Black people (compared with White people) after predicting their IAT scores and completing IATs. One potential interpretation of these findings is that participants simply became less concerned about openly expressing their thoughts and feelings about minority group members, but they may not necessarily think of these openly expressed judgments as being biased. A different interpretation can be derived from research inspired by Monteith and Mark's (2005) theory of prejudice reduction. In a series of studies by Monteith and colleagues (e.g., Czopp, Monteith, & Mark, 2006; Monteith, Devine, & Zuwerink, 1993; Monteith et al., 2001), participants showed enhanced negative self-related affect and increased motivation to control prejudiced behavior when they were confronted with feelings, thoughts, or behaviors that were more prejudiced than their personal standards would permit. From this perspective, the obtained increase in explicit biases and increased alignment between implicit and explicit preferences may reflect increased acknowledgment of bias in the sense that participants noticed a level of bias that conflicts with their personal standards.[1] However, in the absence of additional evidence, any such interpretation may be deemed speculative. Thus, to provide more compelling evidence for the proposed interpretation in terms of acknowledgment of bias, we conducted two sets of studies with different criterion measures. The first set of studies investigated changes in the size of explicit biases and their alignment with implicit biases; the second set of studies measured acknowledgment of bias more directly by asking participants to rate the extent to which they harbor racial biases.

The Present Research

In Studies 1–3, we tested whether predicting one's IAT scores and completing IATs without feedback (Studies 1–3) and with feedback (Study 3) influences explicit preferences for Whites over racial minorities. Using a pre-post repeated-measures design, we tested whether participants would change their explicit preferences from before to after our experimental manipulations such that they show (a) greater alignment between their explicit and implicit preferences and (b) greater explicit preference for White people over racial minorities.

To disambiguate the meaning of the obtained effects on explicit biases, participants in Studies 4–6 were directly asked to assess their level of racial bias. In these studies, participants predicted and/or completed a Black–White IAT without feedback (Studies 4 and 5) or with feedback (Study 5) and then rated the extent to which they harbor racial biases. Study 4 additionally investigated whether the obtained effects are moderated by the degree to which participants hold nonprejudicial goals. Study 6 compared the effects of IAT score prediction with the effects of enhanced attention to one's spontaneous affective reactions toward Black and White individuals. For all studies reported here, the data were collected in one shot without prior statistical analyses. We report all data exclusions, all measures, and all manipulations. All materials, data, and analysis files are available at osf.io/mkc9r.

Study 1

Expanding on the procedures by Hahn et al. (2014), participants in Study 1 provided feeling thermometer ratings of Asians, Blacks, Latinos/Hispanics, and Whites, celebrities and regular people, as well as children and adults at two time points. Between those two times of measurement, we implemented two experimental manipulations. First, after completing the first set of thermometer ratings, half of the participants predicted how they would score on a psychological test designed to measure their "implicit attitudes" toward the same groups; the other half did not make any such predictions. Second and orthogonal to the prediction manipulation, half of the participants completed the corresponding IATs (without feedback), whereas the other half completed the IATs later in the study (see Figure 1). Afterward, all participants completed the same feeling thermometer ratings a second time. Finally, participants completed a set of exploratory measures reported in online supplemental materials Section A. The main question was whether IAT score prediction or actual IAT completion (or both) increase explicit biases and their relation to implicit biases. Although we were primarily interested in bias against minority groups, we included the children–adult and celebrity–regular comparisons to have sufficient variance in each participant's IAT scores to investigate within-subject relations of these scores with participants' prediction scores (and to replicate the findings by Hahn et al., 2014).

Method

Participants and design. One-hundred and 57 undergraduate students from a large Canadian university participated in Study 1 for course credit or payment of $10 CAD. Seven participants responded within less than 300 ms to more than 10% of trials on one or more of the IATs. In line with recommendations by Greenwald, Nosek, and Banaji (2003), these participants were excluded from analyses. Of the remaining 150 participants (65.3% female, median age = 20 years, age range = 18–66 years), 38.7% identified as White, 35.3% as East Asian, 6.7% as South Asian, 5.3%

[1] In the literature on implicit bias, the term *awareness* is often used to refer to people's conscious access to the attitudinal representations that presumably underlie their implicit evaluations (Hahn & Gawronski, 2014; Hahn et al., 2014). This interpretation is different from the one in Monteith and colleagues' work (e.g., Monteith et al., 1993; Monteith & Mark, 2005; Monteith et al., 2001), which is concerned with people's awareness that they are more prejudiced than their personal standards would permit. Although the current research builds on several ideas by Monteith and colleagues, we use the term *acknowledgment of bias* in the current article when we refer to the construct of admitting to a bias, to avoid potential confusion between the two meanings of the term *awareness*.

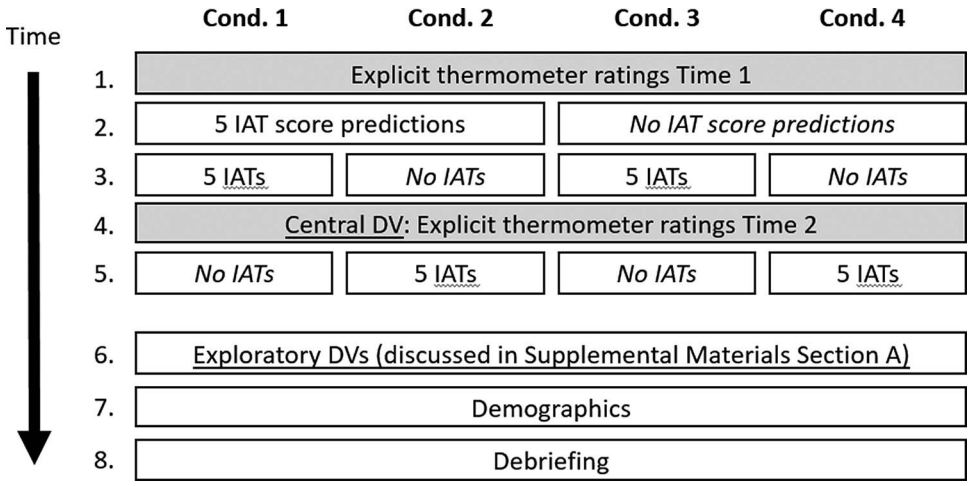

Figure 1. Procedure and design of Study 1: Four between-subjects conditions and one repeated-measures factor in a 2 (IAT Score Prediction: prediction vs. no prediction; between-subjects) × 2 (IAT Completion: before Time 1 thermometer ratings vs. after Time 2 thermometer ratings; between-subjects) × 2 (Time of Feeling Thermometer Ratings: Time 1 vs. Time 2; within-subjects) mixed design.

as Middle Eastern, 2.7% as Black, and 1.7% as Latino. The remaining 10% identified as "other," reporting either a mix of several or a different ethnic background from those mentioned above.

To investigate whether participants would change their explicit evaluations to be more in line with the patterns of their implicit evaluations, we employed a multilevel design in which participants were presented with five pairs of social groups. This multilevel design was adapted from Hahn et al. (2014) with the exception that we manipulated whether participants (a) predicted their IAT scores and (b) completed IATs before they provided their second explicit ratings of the groups (see Figure 1). Thus, in addition to the continuous multilevel aspect of the design (which we explain in more detail below), the study consisted of a 2 (Time of Feeling Thermometer Ratings: Time 1 vs. Time 2) × 2 (IAT Score Prediction: prediction vs. no prediction) × 2 (IAT Completion: completed vs. not completed) mixed design, with the first variable being a within-subjects factor and the other two being between-subjects factors.

Measures and materials. Explicit evaluations were measured with feeling thermometer scales. Participants were asked to indicate their feelings toward eight groups on scales ranging from 0 (*unfavorable and cool feelings*) to 100 (*favorable and warm feelings*). The feeling thermometer items were presented in three blocks: (a) Asians, Blacks, Latinos/Hispanics, and Whites, presented in individually randomized orders; (b) children and adults; and (c) celebrities and regular people (noncelebrities). The presentation order of the three blocks was randomized for each participant. At Time 2, after completing the experimental procedures in the different conditions, participants were asked to *rate your feelings toward the social groups one more time, using the feeling thermometer again*.

Participants in the prediction condition were presented with the following instructions:

> Knowing Your Implicit Attitude: In this study, we are interested in divergences that might occur between people's IMPLICIT and their EXPLICIT attitudes. For this purpose, this study uses a method called the "IMPLICIT ASSOCIATION TEST," or IAT for short. In a minute, you will complete some IATs and we are interested in whether you can predict your performance on each one. That is, we are interested in whether YOU KNOW your IMPLICIT attitude.

On the next slide, participants were told that they would be asked to complete a BLACK–WHITE IAT, a LATINO–WHITE IAT, an ASIAN–WHITE IAT, a CHILD–ADULT IAT, and a CELEBRITY–REGULAR PERSON IAT. Participants then proceeded to making predictions for each IAT, one at a time in individually randomized orders. The prediction slides showed the 10 pictures for each group that were used as target stimuli in the IATs and a short text that referred to the depicted individuals as representing the two social categories in question. Below this text, a sentence read *I predict that the IAT comparing my reactions to* [ASIAN/BLACK/LATINO/CHILD/CELEBRITY, respectively] versus [WHITE/ADULT/REGULAR, respectively] will show that my implicit attitude is . . . and participants were asked to indicate their response on 11-point scales ranging from −5 (*a lot more positive toward ASIAN/BLACK/LATINO/CHILD/CELEBRITY*) to 5 (*a lot more positive toward WHITE/ADULT/REGULAR*).

Participants in the no-prediction condition completed a filler task titled *Consumer Preferences*. These participants were asked to indicate their preferences regarding five types of consumer goods (i.e., foods, cellphone vs. landline, types of movies, leisure time activities, and formal vs. casual clothing) using scales that were similar to the ones in the prediction condition, but without reference to social groups or IATs.

Participants completed the same five IATs described in Hahn et al. (2014), which are shorter than conventional IATs to avoid fatigue. Before beginning with the actual IATs, participants completed a 20-trial positive and negative word-sorting task. They then completed the five shortened IATs in individually randomized orders without repeating the word sorting block (positive and negative words were sorted with the same key assignments in all combined blocks of all IATs). The IATs were introduced with the following two sentences: *You will now complete a group of tasks known as the "IAT."*

These tasks involve CATEGORY JUDGMENTS. The introduction of the IATs was followed by procedural information about the task and keys involved. Each IAT consisted of four blocks: (a) a 20-trial practice block in which participants were asked to categorized faces as WHITE, ADULT, or REGULAR using a right-hand key and BLACK, ASIAN, LATINO, CHILD, or CELEBRITY using a left-hand key; (b) a 40-trial compatible dual-categorization block in which WHITE, ADULT, or REGULAR were grouped with positive words and BLACK, ASIAN, LATINO, CHILD, or CELEBRITY were grouped with negative words; (c) another 40-trial practice block where the pictures had to be categorized with a reversed key assignment; and (d) a 40-trial incompatible dual-categorization block in which BLACK, ASIAN, LATINO, CHILD, or CELEBRITY were grouped with positive words and WHITE, ADULT, or REGULAR were grouped with negative words. IAT scores were computed by comparing the average response latencies of Blocks 4 and 2 divided by their pooled standard deviation (see Greenwald et al., 2003). Thus, higher scores reflect greater bias in favor of WHITE, ADULT, or REGULAR compared with the respective contrast categories.

Each group was represented by photographs of five male and five female faces with neutral expressions, presented with hair and neck against gray backgrounds. The pictures were adapted from Hahn et al. (2014), who used images from the Productive Aging Lab website (Minear & Park, 2004) and images that were publicly available online. The 10 pictures used to represent the category WHITE were different in each of the IATs. Participants did not receive any feedback on their IAT scores in this study. Despite the lower number of trials compared with standard IATs, all five IATs showed satisfactory reliabilities (Cronbach's alpha values for BLACK–WHITE = .73, ASIAN–WHITE = .73, LATINO–WHITE = .72, CHILD–ADULT = .64, CELEBRITY–REGULAR = .68).[2]

Procedure. The procedure of Study 1 is depicted graphically in Figure 1. After providing informed consent, participants started the study by completing the Time 1 thermometer ratings. Afterward, participants in the *prediction* condition predicted their IAT scores as described above, whereas participants in the *no-prediction* condition completed the filler task. Next, participants in the *completion* condition completed the five IATs and then provided their Time 2 thermometer ratings; participants in the *no-completion* condition provided their Time 2 thermometer ratings and then completed the five IATs. The Time 2 thermometer ratings were the same in all conditions and did not refer to the predictions or the IATs. The prediction task emphasized that an IAT score reflects a separate construct from a "feeling" toward a group of people captured by an explicit thermometer rating (see above). The five IATs were presented in orders individually randomized for each participant. After all participants had completed the IATs and the Time 2 thermometer ratings, they completed a set of exploratory measures described in online supplemental materials Section A. The study concluded with demographic questions and an opportunity for participants to provide feedback, after which they were debriefed and compensated for their participation.

Results

Prediction accuracy. Although not the primary question of this study, we first investigated the degree of prediction accuracy among the 75 participants who predicted their IAT scores. Toward this end, we regressed person-standardized IAT scores onto each participant's person-standardized predictions on Level 1 of a multilevel analysis, and then tested the fixed effects of the resulting slopes on Level 2 across participants (see Hahn et al., 2014, for a more detailed account of the rationale behind this within-subject assessment of accuracy). This fixed effect slope is equivalent to an average within-subject correlation between prediction scores and IAT scores for each participant, indicating how accurately participants predicted the patterns of their IAT scores on average. The size of this slope was $b = .48$, $SE = .046$, $t(373) = 10.38$, $p < .001$. Computing correlations separately for each participant revealed a skewed distribution with the same mean as the multilevel analysis, and a median correlation of $r = .54$. Both of these values are slightly lower compared with the ones reported in Hahn et al. (2014), who found a mean correlation of $r = .54$ and a median correlation of $r = .68$ across studies. Nevertheless, the current results still indicate a substantial degree of accuracy in the prediction of IAT scores, replicating the findings by Hahn et al. (2014).

IAT scores. To investigate whether IAT scores were affected by our experimental manipulations, we conducted five separate 2 (IAT Score Prediction) × 2 (IAT Completion) ANOVAs, one for each of the five IATs. There were no significant differences in implicit bias levels across conditions on any of the five IATs, all Fs < 2.40, all ps $> .12$. To investigate if the prediction task influenced the rank order of implicit bias scores for the different target groups (e.g., as a result of participants trying to produce IAT scores that are congruent with their predicted scores), we averaged the scores of each of the five IATs separately within the predictions and the no-predictions conditions and calculated the correlation between these average scores on the five IATs in the two conditions. This correlation was $r = .99$, indicating that participants who predicted their IAT scores produced virtually identical patterns of implicit bias as participants who did not predict their scores. Thus, there was no evidence that the prediction task affected IAT scores.

Alignment of explicit and implicit biases. Our main question was whether predicting IAT scores or actual IAT completion (or both) would lead participants to incorporate their implicit preferences to a greater extent into their explicit preferences. In terms of the current design, enhanced incorporation of implicit preferences into explicit preferences is reflected in a change in the relation between implicit and explicit preference scores from Time 1 to Time 2, such that explicit preference scores become more in line with the patterns of IAT scores (see Hahn et al., 2014). Causal effects of IAT score prediction and IAT completion would be reflected in significant interaction effects, such that the predicted increase in the relation between implicit and explicit preference scores depends on either of the two experimental factors (or both).

To test these effects, we computed five difference scores from the thermometer ratings reflecting relative preferences for the respective groups at Time 1 and Time 2, equivalent to those captured by the five IATs: preference for Whites over Asians, preference for Whites over Blacks, preference for Whites over Latinos, preference for adults over children, and preference for celebrities over regular people. We then

[2] Cronbach's α values were calculated by computing four separate IAT *D*-scores for the first, second, third, and fourth sets of 10 trials of the compatible and incompatible blocks, respectively (S. Teige-Mocigemba, personal communication, October 25, 2017).

simultaneously regressed person-standardized values of each participant's thermometer preference scores at Time 2 onto their person-standardized thermometer preference scores at Time 1 and their IAT scores for each participant on Level 1 of our multilevel analysis. The slope of the Time 1 thermometer preference scores captures the degree to which the patterns of explicit preferences remained stable from Time 1 to Time 2. The slope of the IAT scores captures changes in the pattern of explicit preferences from Time 1 to Time 2 that can be explained by participants' IAT scores (i.e., the average additional variance per participant in Time 2 thermometer preference score patterns that is shared with IAT scores over and above Time 1 thermometer preference scores). A significant effect for the latter slope would indicate a change in the observed patterns of explicit preferences in the sense that they become more in line with the observed patterns of implicit preferences at Time 2 compared with Time 1. To investigate whether this effect depends on the prediction of IAT scores and/or actual completion of IATs, we then tested whether the predicted slope for IAT scores at Level 1 is moderated by IAT score prediction and IAT completion at Level 2 in a 2 × 2 design. Toward this end, we included one contrast code for the IAT score prediction factor (coded −1 for the no-prediction condition and 1 for the prediction condition), one contrast code for the IAT completion factor (coded −1 for IATs not yet completed and 1 for IATs completed), and one contrast code for their interaction (i.e., the product of the two contrasts). We then tested whether the Level 1 effects of IAT scores and Time 1 thermometer preferences interacted with the three Level 2 contrasts.

The results of this analysis are presented in Table 1. A significant main effect of Time 1 thermometer preferences, $b = .73$, $SE = .030$, $t(143.84) = 24.18$, $p < .001$, indicated a substantial degree of consistency in explicit preferences over time. There was also a significant main effect of IAT scores over and above Time 1 thermometer preferences, $b = .15$, $SE = .026$, $t(149.67) = 5.59$, $p < .001$, supporting the predicted adaptation of explicit preferences to implicit preferences. Both of these effects interacted with IAT score prediction. The effect of Time 1 thermometer preferences was smaller in the prediction condition compared with the no-prediction condition, $b = -.12$, $SE = .030$, $t(143.84) = -3.92$, $p < .001$. Conversely, the effect of IAT scores was larger in the prediction compared with the no-prediction condition, $b = .13$, $SE = .026$, $t(149.67) = 4.84$, $p < .001$.

Using dummy codes, we decomposed these interactions and calculated simple slopes for the effects of both predictors within the two experimental conditions (see Table 2). When participants predicted their IAT scores, their IAT scores significantly predicted their Time 2 thermometer preferences over and above their Time 1 thermometer preferences (upper two cells). In contrast, when they did not predict their IAT scores, there was no relation between Time 2 thermometer preferences and IAT scores over and above Time 1 thermometer preferences, and Time 2 thermometer preferences were more consistent with Time 1 ratings (lower two cells). That is, participants who predicted their IAT scores adapted their explicit preferences to be more in line with the patterns of their implicit preferences, whereas participants who did not predict their IAT scores did not adapt their explicit preferences.

Another way to illustrate the increased incorporation of implicit preferences into explicit preferences in the prediction as opposed to the no-prediction conditions is to compare the average within-subjects correlations between implicit and explicit preferences before and after the prediction manipulation. At baseline, the average within-subject correlation between Time 1 thermometer preferences and IAT scores was .21 (95% CI [.13, .29]) across conditions, which is similar to the average size of implicit-explicit correlations in published meta-analyses (e.g., Hofmann, Gawronski, Gschwendner, Le, & Schmitt, 2005). When participants did not predict their IAT scores, the correlation at Time 2 was similar with $r = .22$ (95% CI [.11, .33]). However, when participants predicted their IAT scores, the correlation at Time 2 was larger with $r = .37$ (95% CI [.26, .48]), demonstrating that participants' patterns of explicit preferences became more in line with their implicit preferences.[3]

Whether or not participants completed IATs had no significant effect on whether they adapted their explicit preferences to their implicit preferences, as indicated by a nonsignificant interaction between IAT scores and IAT completion, $b = .03$, $SE = .026$, $t(149.67) = 1.17$, $p = .25$. Although the simple slopes in Table 2 might suggest that participants in the prediction condition adapted

Table 1
Feeling Thermometer Preferences at Time 2 Regressed Simultaneously Onto IAT Scores and Feeling Thermometer Preferences at Time 1 (Level 1) as a Function of IAT Score Prediction and IAT Completion (Level 2), Study 1

Parameters (DV: Thermometer ratings Time 2)	Slope estimates	(Standard errors)
Fixed effects		
Thermometer ratings Time 1	.73***	(.030)
Therm. Time 1 × IAT Completion	−.05†	(.030)
Therm. Time 1 × Prediction	−.12***	(.030)
Therm. Time 1 × Prediction × IAT Completion	.01	(.030)
IAT scores	.15***	(.026)
IAT Scores × IAT Completion	.03	(.026)
IAT Scores × Prediction	.13***	(.026)
IAT Scores × Prediction × IAT Completion	.01	(.026)
Random effect variances		
Thermometer ratings Time 1	.075***	(.017)
IAT scores	.037**	(.012)
Residuals	.211***	(.014)
Goodness of fit		
−2 log likelihood	1198.80	

Note. Level 1 variables and IAT scores were standardized for each individual participant before they were entered in the analysis. Hence, all individual mean values are 0 and no intercept and no Level 1 main effects can be estimated. The prediction manipulation is coded "−1" for no predictions and "1" for predictions; IAT completion is coded "−" when no IATs completed before Time 2 ratings, and "1" when all IATs were completed before Time 2 ratings.
† $p = .087$. ** $p < .01$. *** $p < .001$.

[3] Testing whether these average within-subjects correlations are significantly different from each other involves the same analysis described earlier in this section, but without controlling for Time 1 thermometer preference ratings (which leads to a reduction in power to detect an effect). Specifically, we regressed person-standardized Time 2 thermometer preferences on IAT scores on Level 1, and then tested whether this relation interacts with condition. Consistent with the overall analysis, this less powerful analysis revealed an interaction between IAT scores and IAT score prediction in predicting Time 2 thermometer preference scores, $b = .08$, $SE = .038$, $t(146.00) = 1.97$, $p = .050$, confirming that Time 2 thermometer preferences were more strongly correlated with IAT scores in the prediction as opposed to the no-prediction condition.

Table 2
Simple Slope Estimates and Standard Errors of Average Simultaneous Within-Subject Effect of IAT Scores and Time 1 Feeling Thermometer Preferences in the Prediction of Time 2 Feeling Thermometer Preferences as a Function of IAT Score Prediction and IAT Completion (N = 150), Study 1

Level-1 predictors by condition	IATs completed	No IATs completed
IAT score prediction		
Thermometer preferences Time 1	.57*** (.061)	.66*** (.060)
IAT scores	.31*** (.053)	.23*** (.051)
No IAT score prediction		
Thermometer preferences Time 1	.79*** (.062)	.91*** (.060)
IAT scores	.04 (.053)	.00 (.052)

Note. Scores were standardized for each participant before they were entered into the multilevel analyses. Values can be interpreted similarly to semipartial correlation coefficients.
*** $p < .001$.

their Time 2 thermometer preferences more when they also completed IATs than when they did not complete IATs (.31 as opposed to .23, upper two cells), the difference between the two slopes was not statistically significant, $b = .04$, $SE = .037$, $t(147.17) = 1.12$, $p = .27$. The three-way interaction between the IAT scores, IAT score prediction, and IAT completion was not statistically significant either, $b = .01$, $SE = .026$, $t = 0.40$, $p = .69$.

In addition to these findings, there was a nonsignificant trend for participants to report less consistent thermometer preferences when they did than when they did not complete IATs, reflected in a marginal interaction between thermometer preferences at Time 1 and IAT completion, $b = -.05$, $SE = .030$, $t(143.84) = -1.72$, $p = .087$ (see Table 1, cells on the left vs. right). This effect could be due to the larger time difference between the two sets of thermometer ratings when participants completed IATs between them.

Because the present study included a very diverse sample, with only 38.7% identifying exclusively as White (see above), it is worth noting that none of the effects described above interacted with the ethnic background of the participants. The adaptation of explicit preferences to implicit preferences was similar when we split the sample by ethnic background. The interaction of IAT score prediction and IAT scores as well as the interaction of IAT score prediction and Time 1 thermometer preferences were not qualified by significant three-way interactions with a White versus non-White contrast, both ts < 1, both ps $> .34$.

In sum, IAT score prediction led to a change in explicit preferences from Time 1 to Time 2, such that participants showed a pattern of explicit preferences at Time 2 that was more in line with their implicit preferences. This result is especially remarkable given that participants did not receive any feedback about their IAT scores and half of them had not yet completed any IATs. Actual completion of the IATs did not lead to any adaptation of explicit preferences to implicit preferences, nor did it significantly amplify the effect of predicting IAT scores.

Explicit pro-White bias. Following Hahn et al. (2014), we also tested whether the increased alignment between implicit and explicit preferences is associated with greater explicit pro-White bias. Such an effect may indicate that the IAT score prediction task makes participants realize that they are more biased than they would like, which has been claimed to be a necessary first step in counteracting discrimination (e.g., Monteith & Mark, 2005). Note that an increase in mean-levels does not necessarily follow from increased implicit-explicit correlations. Increased implicit-explicit correlations may occur without participants reporting more bias (e.g., when increased correlations result from increased explicit bias against some groups decreased explicit bias against others); and participants may report more bias even if implicit-explicit correlations remain unchanged (e.g., when participants show the same increase in explicit bias for all groups). Hence, we tested whether predicting IAT scores increases the overall level of explicit bias against minorities in a separate statistical analysis.[4]

Toward this end, we first averaged explicit pro-White preferences across the three minority groups (i.e., Asian vs. White, Black vs. White, and Latino vs. White; Cronbach's alpha Time 1 = .64, Cronbach's alpha Time 2 = .74). We then submitted these average explicit pro-White preference scores to a 2 (Time of Feeling Thermometer Ratings) × 2 (IAT Score Prediction) × 2 (IAT Completion) mixed ANOVA with repeated measures on the first factor. Results are presented in Figure 2. To simplify the interpretation of the figure, the first factor is presented as a change score, reflecting the difference between thermometer preference scores at Time 2 and Time 1. Positive values on this score indicate that participants reported greater explicit pro-White preference at Time 2 compared with Time 1; negative values indicate that they reported smaller explicit pro-White preference at Time 2 compared with Time 1.

There was a significant two-way interaction of IAT score prediction and time, $F(1, 146) = 12.48$, $p < .001$, $\eta_p^2 = .079$. Participants who predicted their IAT scores showed greater explicit bias against minorities at Time 2 compared with Time 1, $F(1, 146) = 9.25$, $p = .003$, $\eta_p^2 = .060$, whereas participants who did not predict their IAT scores, showed marginally smaller explicit bias at Time 2 compared with Time 1, $F(1, 146) = 3.82$, $p = .053$, $\eta_p^2 = .025$. IAT completion did not have any effects on changes in explicit bias from Time 1 to Time 2. The two-way interaction of IAT completion and time as well as the three-way interaction of IAT score prediction, IAT completion, and time were not statistically significant, both Fs < 1, both ps $> .55$.

To investigate whether ratings of Whites or ratings of minorities were primarily responsible for this change, we examined changes in those ratings separately (Cronbach's alpha for the average absolute ratings of the three minority groups Time 1 = .71, Time 2 = .79). Results showed that the obtained increase in explicit pro-White bias stems mainly from the fact that participants reported less warmth toward the three minority groups after as opposed to before predicting their IAT scores ($M_{\text{Change}} = -2.284$, $SE = 1.10$), $F(1, 146) = 4.28$, $p = .040$, $\eta_p^2 = .028$, whereas average warmth ratings for minorities in the no-prediction condi-

[4] We did not test for bias level changes for children or celebrities, because we had no hypotheses about the direction of such changes. Acknowledging the reactions reflected in implicit evaluations of children and celebrities may lead some participants to report more negative and others to report more positive evaluations towards these groups. Hence, we consider these results less relevant for our main question regarding bias against minorities. All mean level ratings of all individual groups at both Time 1 and Time 2 are presented in online supplemental materials Section B.

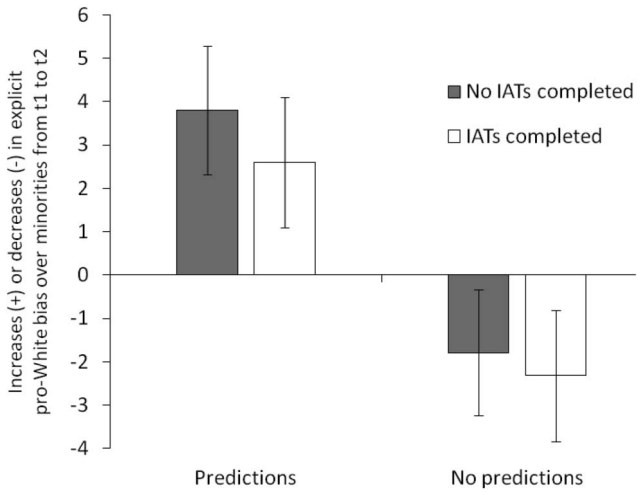

Figure 2. Changes in explicit pro-White bias over Asians, Blacks, and Latinos from Time 1 to Time 2 as a function of IAT score prediction and IAT completion, Study 1. Error bars represent standard errors of predicted values calculated from a 2 (Time) × 2 (IAT Score Prediction) × 2 (IAT Completion) × 3 (Target Group) ANOVA.

tion did not significantly differ before and after the prediction task (M_{Change} = 1.67, SE = 1.10), $F(1, 146) = 2.30$, $p = .132$, $\eta_p^2 = .015$. Ratings of White targets did not change in response to the manipulations, all Fs < 1.1., all ps > .30 (individual ratings are presented in online supplemental materials Section B).

As with the adaptation results, the effect of IAT score prediction on mean-level changes in explicit pro-White bias over time did not interact with participants' own minority status, $F(1, 142) = 0.00$, $p = .99$, $\eta_p^2 = .000$. Participants of all backgrounds reported more explicit pro-White bias (or less explicit anti-White bias) when they predicted their IAT scores than when they did not. Non-White participants reported less pro-White bias (and in some cases anti-White bias) overall, which was reflected in a significant main effect of ethnic background on thermometer preference scores, $F(1, 142) = 5.11$, $p = .025$, $\eta_p^2 = .035$. However, the observed changes in explicit preference scores from Time 1 to Time 2 as a result of IAT score prediction were in the same direction regardless of the minority status of the participants.

Discussion

Hahn et al. (2014) found that participants changed their explicit preferences to be more in line with their implicit preferences and reported more explicit pro-White bias after predicting and completing IATs. The goal of Study 1 was to test whether this effect that was due to IAT score prediction, IAT completion, or a combination of both. Results showed that predicting IAT scores was sufficient to produce change in explicit preferences. Participants who predicted their scores on future IATs subsequently incorporated their implicit preferences more into their explicit preferences, and this effect was associated with greater mean levels of explicit pro-White bias. Completing IATs did not have any such effects by itself, and it did not significantly interact with IAT score prediction. Together, these results are consistent with the idea that acknowledgment of bias can be increased by directing people's attention to their spontaneous affective reactions (e.g., by asking them to predict their scores on future IATs). Yet, they question the idea that IAT completion increases acknowledgment of bias by giving people an opportunity to observe behavioral effects of their attitudes. They are also inconsistent with the notion that mere anticipation of a bias test increases acknowledgment of bias. Although the current study did not include a separate "mere anticipation" condition, the finding that IAT completion alone did not lead to more acknowledgment than the control condition (where there was no announcement of a test at any point before IAT completion) speaks against the interpretation that anticipation of a bias test increased acknowledgment of bias in the current study.

Study 2

One potential reason for the null effect of IAT completion in Study 1 is that participants did not realize that the IATs assessed their implicit biases, given that the IAT was introduced as a "categorization task" rather than a measure of implicit bias. Because each IAT was named with the labels of the social groups (e.g., "BLACK–WHITE IAT") and these names were repeatedly mentioned in the instructions, we find such an interpretation rather implausible. Still, to ensure participants understood that the IAT is supposed to be a measure of implicit bias, we included this information in the instructions of Study 2, which aimed to replicate the findings of Study 1 with a sample of German university students.

Germany has been one of the most immigrated-to countries in the world over the past 50 years, with 13.5% of the population reporting being born outside of Germany and an additional 6.8% reporting non-German family backgrounds in 2014 (German Federal Agency for Civic Education, 2016; German Federal Statistical Office, 2015). Hence, issues of discrimination and prejudice against ethnic minorities are at the center of heated societal debates. However, because the largest groups of immigrants in Germany have come from Eastern and South Eastern Europe, Turkey, and the Middle East (German Federal Statistical Office, 2015), the social categories around which these discussions revolve align more along cultural differences, religion, and ethnicity, than along the racial categories that are typically at the center of North American debates.[5] Thus, a replication in Germany, where ethnic discrimination and prejudice are salient discussion topics, but participants can be assumed to have different experiences with the specific groups in question, represents an interesting extension to Study 1 (see Hahn, Judd, & Park, 2010, for more details on national differences and diversity).

Method

Participants and design. We aimed at recruiting approximately 200 participants. Using the effect sizes obtained in Study 1 and the GPower software to estimate statistical power (Faul, Erdfelder, Lang, & Buchner, 2007), a sample of 200 provides a power of 97% in replicating the obtained adaptation of explicit prefer-

[5] In 2014, the proportion of the German population with backgrounds in Africa, Asia, or Latin America, was 0.7%, 1.0%, and 0.3%, respectively (German Federal Statistical Office, 2015).

ences to implicit preferences.[6] Two-hundred and five participants from a large urban University in Germany completed the study in exchange for experimental credit or a payment of 6€ (and some candy). Ten of these participants indicated having participated in an earlier study with the same IAT prediction paradigm. Two additional participants responded within less than 300 ms on more than 10% of the trials on one or more of the IATs (see Greenwald et al., 2003). Data from these participants were excluded from the following analyses. Of the remaining 193 participants (81.9% female, median age = 22 years, age range = 17–44 years), 70.5% reported exclusive German ancestry, 23.3% indicated being born in Germany from one or two parents with non-German backgrounds, and 8.8% indicated being foreign-born themselves. When asked about racial categories (relevant for the IATs in the study), 84.5% of the sample identified exclusively as White, 5.2% as Middle Eastern or both White and Middle Eastern, 3.1% as East Asian, 2.1% as both White and Latino, and 1.0% as South Asian. The remaining 4.1% indicated "other" or did not report any ethnic background.[7] In addition to the continuous multilevel aspect of the design (see Study 1), the study consisted of a 2 (Time of Feeling Thermometer Ratings: Time 1 vs. Time 2) × 2 (IAT Score Prediction: prediction vs. no prediction) × 2 (IAT Completion: completed vs. not completed) mixed design, with the first variable being a within-subjects factor and the other two being between-subjects factors.

Materials and procedure. The materials and procedures were similar to Study 1, except for a few minor changes in the procedure and the instructions. All materials were translated and administered in German, which led to minor changes in the wording of the prediction task. IAT score predictions were measured with 7-point instead of 11-point scales ranging from 1 (*substantially more positive toward ASIAN* [BLACK/LATINO/CELEBRITY/CHILD, respectively]) to 7 (*substantially more positive toward WHITE* [REGULAR/ADULT, respectively]). Participants also read one additional paragraph before making their predictions, detailing that philosophers, anthropologists, and psychologists have long suspected that there are different kinds of attitudes. The purpose of this paragraph was to emphasize that implicit preferences reflect a construct that is different from explicit preferences, which should reduce potential demand to align explicit with implicit preferences.

Participants in Study 2 also received more information about the IAT. The additional information read: *You will now complete a series of tasks known as the IAT—the implicit association test. The IAT measures your implicit attitudes*. As in Study 1, each individual IAT was introduced with the social categories in question (e.g., *the next four tasks belong to the BLACK–WHITE IAT*). The IATs again showed satisfactory reliabilities (Cronbach's alpha values for BLACK–WHITE = .78, ASIAN–WHITE = .68, LATINO–WHITE = .74, CHILD–ADULT = .61, CELEBRITY–REGULAR = .59).

After signing informed consent, participants completed the first thermometer ratings for the eight groups described in Study 1. Next, participants in the prediction condition predicted their scores on the five IATs on the 7-point scales described above, while participants in the no-prediction condition answered the five consumer preference questions from Study 1 on similar 7-point scales. Afterward, half of the participants completed the IATs and then provided their second thermometer ratings, while the other half went straight to completing the second thermometer ratings. Next, all participants completed the exploratory measures described in online supplemental materials Section A. Participants in the no-completion condition were asked to complete the five IATs after the exploratory measures. Finally, all participants provided demographic information, including information about previous participation in the study and optional feedback, before they were debriefed and compensated.

Results

Prediction accuracy. Accuracy in the prediction of IAT scores was analyzed in line with the procedures of Study 1. The 95 participants who predicted their IAT scores showed similar levels of accuracy as in Study 1, $b = .49$, $SE = .040$, $t(474) = 12.19$, $p < .001$, median within-subject correlation of $r = .60$.

IAT scores. Separate 2 (IAT Score Prediction) × 2 (IAT Completion) ANOVAs for each of the five IATs revealed that average levels of implicit preferences were unaffected by our manipulations, all Fs < 2.80, all $p > .09$, the only exception being the celebrity-regular person IAT.[8] The correlation between average IAT scores in the prediction and the no-prediction conditions was $r = .97$, again indicating identical patterns of implicit preferences across the two conditions. Together, these findings suggest that the prediction task did not influence IAT scores.

Alignment between explicit and implicit biases. As in Study 1, we conducted a multilevel analysis to test whether (a) participants changed their explicit preferences from Time 1 to Time 2 to be more in line with the patterns of their implicit preferences, and (b) whether this adaptation effect depended on our manipulations of IAT score prediction and IAT completion (or both). Toward this end, we simultaneously regressed participants' Time 2 thermometer preference scores onto their Time 1 thermometer preference scores and their IAT scores on Level 1 (all person-standardized), and then modeled the average per-participant relationships as a function of the two experimental factors and their interactions on Level 2. The results of this analysis are summarized in Table 3; simple slopes in the four experimental conditions are depicted in Table 4.

Results showed a significant effect of Time 1 thermometer preferences, $b = .80$, $SE = .021$, $t(201.89) = 37.89$, $p < .001$,

[6] Power estimates for the multilevel analyses were obtained by running individual regressions for each participant, saving the resulting slopes per participant, and then running ANOVAs on those slopes and calculating η_p^2-values. By ignoring the multilevel nature of the calculations, they likely underestimate the actual statistical power in replicating the obtained effects.

[7] The proportions of ethnic minority backgrounds in the four experimental conditions were too small to allow for statistical comparisons between participants with different backgrounds in this study.

[8] For the celebrity–regular IAT, we found a significant main effect of IAT score prediction, $F(1, 189) = 6.67$, $p = .011$, $\eta_p^2 = .034$; a marginal main effect of IAT completion, $F(1, 189) = 2.88$, $p = .092$, $\eta_p^2 = .015$; and a marginal interaction of the two factors, $F(1, 189) = 3.49$, $p = .063$, $\eta_p^2 = .018$. Inspection of the mean patterns revealed that participants in the prediction plus IAT-at-end-of-study condition showed no average bias with an IAT D score of $D = .053$, $SE = .054$, whereas participants in the other three conditions showed the expected pro-celebrity bias with scores between −.135 and −.182, all SEs = .053. Because this pattern did not replicate in any of the other studies and the celebrity-regular IAT is not of primary concern to our theoretical question, we treat this finding as a false-positive and do not discuss it further.

indicating a substantial degree of consistency in thermometer preferences over time. There was also a significant main effect of IAT scores over and above Time 1 thermometer preferences, $b = .11$, $SE = .020$, $t(201.96) = 5.58$, $p < .001$, supporting the predicted increase in the alignment of explicit and implicit preferences. Replicating the results of Study 1, both effects were qualified by significant interactions with IAT score prediction. Participants who predicted their IAT scores showed less consistency in their thermometer preferences over time, as indicated by a significant interaction of Time 1 thermometer preferences and IAT score prediction, $b = -.05$, $SE = .040$, $t(201.89) = -2.22$, $p = .027$. At the same time, IAT scores showed a stronger relation to Time 2 thermometer preferences for participants who predicted their IAT scores compared with participants who did not predict their IAT scores, indicated by a significant interaction of IAT scores and IAT score prediction, $b = .09$, $SE = .020$, $t(201.96) = 4.51$, $p < .001$. Thus, replicating the finding of Study 1, prediction of IAT scores led participants to adapt the patterns of their explicit preferences to be more in line with their implicit preferences.

Analyses of simple within-subject implicit-explicit correlations showed that IAT scores and thermometer preferences were correlated with $r = .32$ (95% CI [.26, .38]) at Time 1. At Time 2, this correlation increased to $r = .45$ (95% CI [.37, .53]) for participants who predicted their IAT scores, but it remained at $r = .30$ (95% CI [.22, .38]) for participants who did not predict their IAT scores.[9]

In addition to the effects of IAT score prediction, there was a marginal interaction between IAT scores and IAT completion, $b = .04$, $SE = .020$, $t(201.96) = 1.88$, $p = .062$, indicating that IAT scores showed a stronger relation to Time 2 thermometer preferences for participants who had completed the IATs by the time they provided their second thermometer scores compared with participants who had not completed the IATs (compare slopes for IAT scores in cells on the left with cells on the right-hand side in Table 4). Note, however, that (a) the effect of IAT completion alone was much weaker than the effect of IAT score prediction alone, and only marginally significant in itself (adaptation slope of .07 compared with an adaptation slope of .17); (b) the critical interaction effect was only marginally significant; and (c) no such effect was found in Study 1.

The interaction of Time 1 thermometer preferences and IAT score prediction was further qualified by a significant three-way interaction with IAT completion, $b = .04$, $SE = .021$, $t(201.89) = 2.12$, $p = .035$. As shown in Table 2, participants who had neither completed the IATs nor predicted their IAT scores showed almost identical patterns in their thermometer preferences at Time 1 and Time 2 (slope based on standardized values at .93, lower right cell), whereas consistency in thermometer preferences was lower in the other three conditions (each of which included at least one intervention). This result suggests that either one of the two interventions (i.e., IAT score prediction, IAT completion) was sufficient to reduce the consistency of thermometer preferences from Time 1 to Time 2 in Study 2 (a minor variation on the results of Study 1, where IAT score prediction and IAT completion each independently lowered consistency in two independent two-way interactions).

Explicit pro-White bias. To investigate changes in mean levels of explicit bias against minorities, we again averaged explicit pro-White bias across the three minority groups (Cronbach's alpha Time 1 = .63, Cronbach's alpha Time 2 = .73) and submitted these preference scores to the same 2 (Time of Feeling Thermometer Ratings) × 2 (IAT Score Prediction) × 2 (IAT Completion) mixed ANOVA with repeated measures on the first factor. Results are presented in Figure 3, depicting average change

Table 3
Feeling Thermometer Preferences at Time 2 Regressed Simultaneously Onto IAT Scores and Feeling Thermometer Preferences at Time 1 (Level 1) as a Function of IAT Score Prediction and IAT Completion (Level 2), Study 2

Parameters (DV: Thermometer ratings Time 2)	Slope estimates	(Standard errors)
Fixed effects		
Thermometer ratings Time 1	.80***	(.021)
Therm. Time 1 × IAT Completion	−.03	(.021)
Therm. Time 1 × Prediction	−.05*	(.021)
Therm. Time 1 × Prediction × IAT Completion	.04*	(.021)
IAT scores	.11***	(.020)
IAT Scores × IAT completion	.04†	(.020)
IAT Scores × Prediction	.09***	(.020)
IAT Scores × Prediction × IAT Completion	−.01	(.020)
Random effect variances		
Thermometer ratings Time 1	.028**	(.009)
IAT scores	.023**	(.008)
Residuals	.182***	(.011)
Goodness of fit		
−2 log likelihood	1296.98	

Note. Level 1 variables and IAT scores were standardized for each individual participant before they were entered in the analysis. Hence, all individual mean values are 0 and no intercept and no Level 1 main effects can be estimated. The prediction manipulation is coded "−" for no predictions and "1" for predictions; IAT completion is coded "−" when no IATs completed before Time 2 ratings, and "1" when all IATs were completed before Time 2 ratings.
† $p = .062$. * $p < .05$. ** $p < .01$. *** $p < .001$.

Table 4
Simple Slope Estimates and Standard Errors of Average Simultaneous Within-Subject Effect of IAT Scores and Time 1 Feeling Thermometer Preferences in the Prediction of Time 2 Feeling Thermometer Preferences as a Function of IAT Score Prediction and IAT Completion (N = 193), Study 2

Level-1 predictors by condition	IATs completed	No IATs completed
IAT score prediction		
Thermometer preferences Time 1	.76*** (.041)	.74*** (.045)
IAT scores	.23*** (.039)	.17*** (.042)
No IAT score prediction		
Thermometer preferences Time 1	.77*** (.041)	.93*** (.042)
IAT scores	.07† (.039)	−.02 (.041)

Note. Scores were standardized for each participant before they were entered into the multilevel analyses. Values can be interpreted similarly to partial correlation coefficients.
† $p = .087$. *** $p < .001$.

[9] As in Study 1, repeating the analysis without controlling for Time 1 thermometer preference scores confirmed that these average correlations were significantly different from each other, $b = .08$, $SE = .029$, $t(189.00) = 2.57$, $p = .011$.

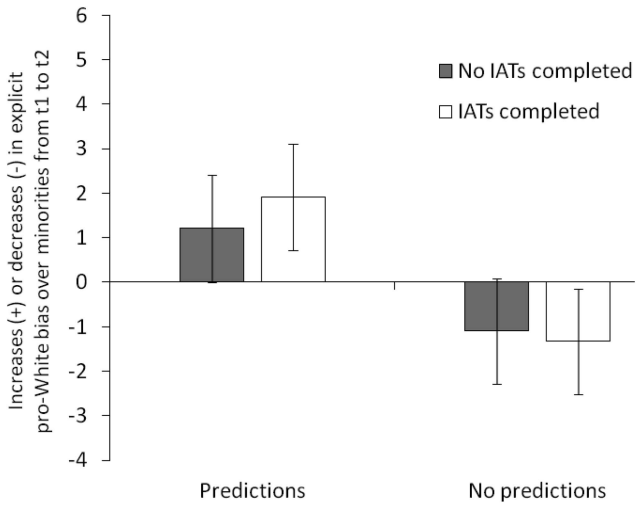

Figure 3. Changes in explicit pro-White bias over Asians, Blacks, and Latinos from Time 1 to Time 2 as a function of IAT score prediction and IAT completion, Study 2. Error bars represent standard errors of predicted values calculated from a 2 (Time) × 2 (IAT Score Prediction) × 2 (IAT Completion) × 3 (Target Group) ANOVA.

in preference scores over time. Replicating the results of Study 1, there was a significant interaction of time and IAT score prediction, $F(1, 189) = 5.38$, $p = .021$, $\eta_p^2 = .028$. Participants who predicted their IAT scores reported marginally more explicit pro-White bias over time, $F(1, 189) = 3.36$, $p = .068$, $\eta_p^2 = .017$, whereas participants who did not predict their IAT scores did not show any changes over time, $F(1, 189) = 2.09$, $p = .15$, $\eta_p^2 = .011$. Again replicating the results of Study 1, there was no significant interaction between IAT completion and time, nor was there a significant three-way interaction of IAT score prediction, IAT completion, and time, both Fs < 0.20, ps $> .65$. Hence, even though participants who completed IATs tended to adapt their explicit preferences to be more in line with their implicit preferences, IAT completion did not influence the overall size of explicit pro-White bias which, if anything, showed an effect in the opposite direction (see bar on the far-right in Figure 3).

Follow-up analyses confirmed that the increase in explicit pro-White bias resulting from IAT score prediction was driven by more negative evaluations of minorities compared with Whites (Cronbach's alpha for absolute minority ratings: Time 1 = .80, Time 2 = .82). Participants' warmth ratings for minorities significantly decreased after IAT score prediction ($M_{\text{Change}} = -2.14$, $SE = .59$), $F(1, 189) = 13.25$, $p < .001$, $\eta_p^2 = .066$, but there was no change in the no-prediction condition ($M_{\text{Change}} = -0.82$, $SE = .58$), $F(1, 189) = 2.01$, $p = .158$, $\eta_p^2 = .011$. Ratings of Whites did not change in response to IAT score prediction, ($M_{\text{Change}} = -0.58$, $SE = .85$), $F(1, 189) = 0.46$, $p = .499$, $\eta_p^2 = .002$, and in fact became more negative in the no-predictions condition ($M_{\text{Change}} = -2.03$, $SE = .84$), $F(1, 189) = 5.87$, $p = .016$, $\eta_p^2 = .030$. Thus, replicating the results of Study 1, increased alignment between explicit and implicit preferences again translated into stronger explicit pro-White (anti-minority) bias (individual ratings are presented in online supplemental materials Section B).

Discussion

Study 2 replicated the main findings of Study 1 in a sample of German participants. As in Study 1, prediction of IAT scores changed explicit preferences such that (a) explicit preferences became more in line with implicit preferences and (b) participants showed greater levels of explicit pro-White bias. These results provide further support for the idea that acknowledgment of bias can be increased by directing people's attention to their spontaneous affective reactions (e.g., by asking them to predict their scores on future IATs).

Different from Study 1, participants in Study 2 also showed a marginal adaptation effect in response to completing IATs. This finding is consistent with the idea that IAT completion may increase acknowledgment of bias by giving participants an opportunity to observe behavioral effects of their attitudes, or by increasing anticipation that one's biases will be revealed (the two factors are not dissociated in Studies 1 and 2). However, no such effect was found in Study 1, the critical interaction effect was only marginal in Study 2, and IAT completion did not increase mean levels of explicit pro-White bias in either study. Thus, even if IAT completion influences acknowledgment of bias via self-perception or via anticipation of measurement, such effects seem to be less reliable and less consistent compared with the effects of IAT score prediction. A potential interpretation of this difference is that many participants interpret their responses on the IAT in a manner that is unrelated to personal bias (Monteith et al., 2001), which may counteract acknowledgment of bias in response to IAT completion.

Study 3

The IAT procedures in the preceding studies differ from the ones in many classroom exercises, online IATs, and bias awareness trainings, in that participants did not receive any feedback on their levels of bias. Thus, a major goal of Study 3 was to investigate whether IAT feedback can increase acknowledgment of bias independent of the obtained effects of IAT score prediction. Although such an effect would be consistent with the idea that people are unable to know their implicit biases without feedback on their personal measurement scores, we deem IAT feedback unlikely to increase acknowledgment of bias for two reasons. First, Hahn et al. (2014) found that people can predict the patterns of their IAT scores with a high level of accuracy (replicated in Studies 1 and 2), which poses a challenge to the idea that people do not know their implicit biases unless they receive personal feedback. Second, research by Howell and colleagues suggests that participants tend to react defensively to IAT feedback when it suggests a level of bias that is stronger than what participants would ascribe to themselves (Howell et al., 2015; Howell & Ratliff, 2017). However, neither of these findings rules out the possibility that IAT feedback increases the alignment of implicit and explicit preferences and overall levels of explicit pro-White bias, as we found for IAT score prediction in Studies 1 and 2. Thus, to address this question more directly, Study 3 used the same design as Study 2, the only difference being the inclusion of an additional condition in the manipulation of IAT completion. Whereas participants in Study 2 did or did not complete IATs without feedback, participants in Study 3 either (a) did not complete IATs, (b) completed the IATs without feedback, or (c) completed the IATs with feedback.

Method

Participants and design. Based on the procedure to estimate statistical power in Study 2, we aimed for a sample size of approximately 240 participants to account for the additional between-subjects conditions.[10] In anticipation of potential exclusions, we recruited 257 participants at the same large urban university in Germany as in Study 2. Participants received either 6€ or experimental credit (and some candy) for their participation. Twelve participants indicated having participated in a study with the same IAT score prediction paradigm before, and two participants responded in less than 300 ms on more than 10% of the IAT trials (see Greenwald et al., 2003). These participants were excluded from the following analyses. Of the remaining 243 participants (79.0% female, median age = 22 years, age range = 18–66 years), 75.6% reported exclusively German ancestry, 18.5% having one or two non-German parents, and 5.8% being foreign-born themselves. Considering racial categories, 86.4% identified as exclusively White, 7.0% Middle Eastern or both White and Middle Eastern, 2.1% as Latino or both White and Latino, one participant (0.4%) identified as Black, and the remaining 4.1% as another category or a mix of several racial categories. In addition to the continuous multilevel aspect of the design (see Study 1), the study consisted of a 2 (Time of Feeling Thermometer Ratings: Time 1 vs. Time 2) × 2 (IAT Score Prediction: prediction vs. no prediction) × 3 (IAT Completion: not completed vs. completed without feedback vs. completed with feedback) mixed design, with the first variable being a within-subjects factor and the other two being between-subjects factors.

IAT feedback. Study 3 contained one additional level in the manipulation of IAT completion: IAT completion with feedback. In this condition, participants completed IATs before their Time 2 thermometer ratings as in the *IAT completed* conditions of Studies 1 and 2, but additionally received feedback on their performance similar to the IAT web page. Using the psychological lab software Inquisit, an automatized script calculated each participant's IAT *D* score (Greenwald et al., 2003) for each of the five IATs and converted these scores into personal feedback statements: *Your data suggest* [. . .] automatic preference for [Group A] over [Group B]. Based on the cutoffs for feedback on the IAT website, the qualifiers in this statement were *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 relevant groups were imputed according to the sign of the *D* score.

Materials and procedure. The materials and procedure were identical to those of Study 2, the only exception being the addition of the IAT completion with feedback condition. After participants provided their Time 1 thermometer ratings, and then predicted (or not) their IAT scores, one third completed the five IATs and received feedback, one third completed the IATs without feedback, and one third did not complete the IATs. Afterward all participants completed the Time 2 thermometer ratings and the exploratory measures of downstream consequences described in online supplemental materials Section A. As in Study 2, participants in the no-completion condition completed the IATs at the end of the study after the exploratory measures. The five IATs again showed satisfactory reliabilities (Cronbach's alpha values for BLACK–WHITE = .71, ASIAN–WHITE = .69, LATINO–WHITE = .69, CHILD–ADULT = .67, CELEBRITY–REGULAR = .57).[11]

Results

Prediction accuracy. Accuracy in the prediction of IAT scores was analyzed in line with procedures of Study 1. The 125 participants who predicted their IAT scores predicted them slightly more accurately than in Studies 1 and 2, with values closer to those reported by Hahn et al. (2014), mean: $b = .51$, $SE = .036$, $t(124) = 14.19$, $p < .001$, median correlation: $r = .64$.

IAT scores. We again found no evidence for mean-level differences in IAT scores across the six conditions, all Fs < 2.10, all $p > .12$. A correlation of $r = .99$ between the patterns of average IAT scores in the prediction and no-prediction conditions further indicated that IAT scores were unaffected by the prediction task.

Alignment between explicit and implicit biases. To test whether participants adapted their explicit preferences to be more line with the patterns of their implicit preferences, we again regressed person-standardized scores of their Time 2 thermometer preference scores onto person-standardized IAT scores and Time 1 thermometer preference scores simultaneously on Level 1 of a multilevel design, and then tested if the resulting slopes differed by condition on Level 2 in a 2 (IAT Score Prediction) × 3 (IAT Completion) design. Because the IAT completion manipulation encompassed three levels in the current study, we used two contrast codes, one that compared the no-completion condition (coded −2) to the two completion conditions (both coded +1), and another contrast code that tested whether the effects of IAT completion with feedback (coded +1) differed from the effects of completing IATs without feedback (coded −1). Model results and simple slopes are presented in Tables 5 and 6.

There was a significant main effect of Time 1 thermometer preferences indicating a substantial level of consistency in thermometer preferences, $b = .78$, $SE = .020$, $t(242.05) = 39.70$, $p < .001$. There was also a significant main effect of IAT scores indicating that participants adapted their Time 2 thermometer preferences to their IAT scores, $b = .14$, $SE = .019$, $t(260.42) = 5.61$, $p < .001$. As in Studies 1 and 2, both effects were qualified by significant interactions with IAT score prediction. Participants who predicted their IAT scores showed significantly less consistency in their thermometer preferences, which was reflected in a significant interaction between Time 1 thermometer preferences and IAT score prediction, $b = -.09$, $SE = .020$, $t(242.05) = -4.71$, $p < .001$. Moreover, participants who predicted their IAT scores adapted their Time 2 thermometer preferences significantly more to their IAT scores, which was reflected in a significant interaction between IAT scores and IAT score prediction, $b = .11$,

[10] GPower estimations suggested a total sample size of 200 participants for Study 3. Based on the current design with six between-subjects conditions, we increased this number to 240 to have at least 40 participants per condition.

[11] There was one difference in the IATs in Study 3 compared with the previous studies. In Studies 1 and 2, participants completed 20 training trials for the sorting of the faces (Block 1) before the initial combined block, and 40 training trials with the reversed sorting (Block 3) before the reversed combined block. In Study 3, the second training block included only 20 trials. Given the satisfactory reliabilities, this difference is not discussed further.

Table 5
Feeling Thermometer Preferences at Time 2 Regressed Simultaneously Onto IAT Scores and Feeling Thermometer Preferences at Time 1 (Level 1) as a Function of IAT Score Prediction and IAT Completion (Level 2), Study 3

Parameters (DV: Thermometer ratings Time 2)	Slope estimates	(Standard errors)
Fixed effects		
Thermometer ratings Time 1	.78***	(.020)
Therm. Time 1 × IAT Compl. Ctr1 (no-IAT vs. IAT)	−.02	(.014)
Therm. Time 1 × IAT Compl. Ctr2 (no fb vs. fb)	−.02	(.024)
Therm. Time 1 × Prediction	−.09***	(.020)
Therm. Time 1 × Prediction × IAT Compl. Ctr1	−.03*	(.014)
Therm. Time 1 × Prediction × IAT Compl. Ctr2	.02	(.024)
IAT scores	.14***	(.019)
IAT Scores × IAT Compl. Ctr1 (no-IAT vs. IAT)	.04**	(.014)
IAT Scores × IAT Compl. Ctr2 (no fb vs. fb)	.02	(.023)
IAT Scores × Prediction	.11***	(.019)
IAT Scores × Prediction × IAT Compl. Ctr1	.01	(.014)
IAT Scores × Prediction × IAT Compl. Ctr2	.01	(.023)
Random effect variances		
Thermometer ratings Time 1	.037***	(.009)
IAT scores	.034***	(.009)
Residuals	.167***	(.008)
Goodness of fit		
−2 log likelihood	1614.16	

Note. All Level 1 variables and the dependent IAT scores were standardized for each individual participant before they are entered in the analysis. Hence, all individual mean values are 0 and no intercept and no Level 1 main effects can be estimated. The prediction manipulation is coded "−" for no predictions and "1" for predictions. The IAT completion manipulation is reflected in two separate contrasts. In Contrast 1 (Ctr1), the no-IAT completion condition is coded "−2" while both IAT completion conditions are coded "1;" in Contrast 2 (Ctr2), the no-completion condition is coded "0," the IAT completion without feedback conditions is coded "−," and the IAT with feedback condition is coded "1." Ctr = Contrast; fb = feedback.

SE = .019, $t(260.42) = 5.61$, $p < .001$. As can be seen from comparing the second to the fourth row of data in Table 6, participants adapted their explicit preferences only when they predicted their IAT scores, but not when they did not predict their IAT scores. In terms of simple within-subjects correlations, Time 1 thermometer preference scores were correlated with IAT scores at $r = .34$ (95% CI [.28, .40]) across conditions. Time 2 thermometer preference scores were correlated with IAT scores at $r = .30$ (95% CI [.22, .38]) in the no-predictions condition, but at $r = .50$ (95% CI [.41, .58]) in the prediction condition.[12]

There was also a significant interaction of IAT scores with the contrast comparing the no-completion condition with the two completion conditions, $b = .04$, *SE* = .014, $t(260.37) = 2.62$, $p = .009$, meaning that IAT completion led to significantly more adaptation than no IAT completion when collapsing across the feedback and no-feedback conditions. Further analyses with additional contrast codes revealed that this effect was driven by the fact that participants adapted their scores more to their IAT scores after they completed IATs with feedback than when they did not complete IATs, $b = .06$, *SE* = .023, $t(265.00) = 2.72$, $p = .007$ (compare slopes of IAT score in left-most column, .33 and .07, with right-most column in Table 6, .16 and −.02). Similar to the effects found in Study 2, the degree of adaptation in the IAT *without* feedback condition (center column) differed only marginally from the no-IAT condition, $b = .04$, *SE* = .023, $t(255.92) = 1.82$, $p = .069$. The IAT completion with or without feedback conditions did not differ significantly from each other, $b = .02$, *SE* = .023, $t(260.44) = 0.89$, $p = .376$.[13] This result suggests that IAT completion with feedback may increase acknowledgment of bias. Note, however, that the effect of IAT plus feedback alone was much smaller than the effect of IAT score prediction (.07 compared with .16), and even though an overall main effect suggested that IAT completion with feedback led to more adaptation than no IAT completion, the simple slope of adaptation in the IAT completion plus feedback condition was not significant in and of itself (lower left-most column), $b = .07$, *SE* = .047, $t(266.21) = 1.47$, $p = .143$. At the same time, none of the possible contrasts we tested for IAT completion interacted with the IAT prediction manipulation, all ts < 1, all ps $> .37$, indicating that the effects of IAT completion were not significantly smaller in the no-prediction compared with the prediction conditions (even if not significant in and of themselves). In sum, the results suggest that IAT completion with feedback may lead to increased alignment of implicit and explicit evaluations overall in addition to other factors, but results remained inconclusive regarding the effect of IAT completion with feedback alone.

There was also a significant three-way interaction between Time 1 thermometer preferences, IAT score prediction, and the contrast

[12] Testing the difference between the two average within-subjects correlations by repeating the described analysis without controlling for Time 1 thermometer scores replicated the significant interaction of IAT scores and IAT score prediction in predicting Time 2 thermometer ratings, $b = .10$, *SE* = .030, $t(237.00) = 3.22$, $p = .001$.

[13] All of these additional contrast codes were coded −1 and +1 for the relevant conditions that are being compared, and include an additional contrast code comparing the third condition (coded −2) with the two conditions in question (both coded +1) in the model for a full set of orthogonal codes.

Table 6
Simple Slope Estimates and Standard Errors of Average Simultaneous Within-Subject Effects of IAT Scores and Time 1 Feeling Thermometer Preferences in the Prediction of Time 2 Feeling Thermometer Preferences as a Function of IAT Score Prediction and IAT Completion (N = 243), Study 3

Level-1 predictors by condition	IATs completed with feedback	IATs completed without feedback	No IATs completed
IAT score prediction			
Thermometer preferences Time 1	.63*** (.047)	.63*** (.048)	.79*** (.047)
IAT scores	.33*** (.047)	.26*** (.046)	.16*** (.046)
No IAT score prediction			
Thermometer preferences Time 1	.84*** (.048)	.91*** (.049)	.85*** (.049)
IAT scores	.07 (.047)	.05 (.048)	−.02 (.047)

Note. Scores were standardized for each participant before they were entered into the multilevel analyses. Values can be interpreted similarly to partial correlation coefficients.
*** $p < .001$.

comparing the two IAT completion conditions with the no IAT completion condition, $b = -.03$, $SE = .014$, $t(241.96) = -2.24$, $p = .026$ (see Table 5). For participants who predicted their IAT scores, thermometer preferences were less consistent when they completed IATs than when they did not complete IATs, $b = -.05$, $SE = .019$, $t(249.55) = -2.73$, $p = .007$ (see first row of data in Table 6: .63 and .63, vs. .79). For participants who did not predict their IAT scores, IAT completion had no effect on consistency in thermometer preferences, $b = .01$, $SE = .020$, $t(234.93) = 0.47$, $p = .636$ (see third row of data in Table 6: .84 and .91 vs. .85). Recall that we found less consistency in thermometer ratings in response to IAT completion in both Studies 1 and 2. The interaction presented here means that in the current study, this effect was replicated only in the prediction condition, but not in the no-prediction condition. However, given that the effect is small and driven by an outlier in the IAT without feedback condition (slope of .91, see Table 6), we believe that this particular pattern of a three-way interaction may be a false positive.

In sum, completing IATs with feedback led to less consistency in thermometer ratings when combined with prediction, and possibly to more adaptation of explicit to implicit preferences, although the latter effect remained inconclusive.

Explicit pro-White bias. We again investigated mean-level changes in explicit pro-White bias by averaging explicit preference for Whites across the three minority groups (Cronbach's alpha Time 1 = .72, Time 2 = .77) and submitting these scores to a 2 (Time of Feeling Thermometer Ratings) × 2 (IAT Score Prediction) × 3 (IAT Completion) mixed ANOVA with repeated measures on the first factor. Figure 4 depicts Time 2 minus Time 1 difference scores as a function of condition. Replicating the results of Studies 1 and 2, there was a significant interaction between time and IAT score prediction, $F(1, 237) = 7.13$, $p = .008$, $\eta_p^2 = .029$. Participants who predicted their IAT scores reported significantly more explicit bias over time, $F(1, 237) = 13.69$, $p < .001$, $\eta_p^2 = .055$, whereas participants who did not predict their IAT scores did not show any changes over time, $F(1, 237) = 0.02$, $p = .90$, $\eta_p^2 = .000$. Also replicating the results of Studies 1 and 2, there was no significant interaction effect of IAT completion and change from Time 1 to Time 2, nor was there a significant interaction effect of IAT score prediction, IAT completion, and change over time, both Fs < 1.50, both ps $> .20$.

Follow-up analyses on individual thermometer ratings (Cronbach's alpha for absolute minority ratings: Time 1 = .76, Time 2 = .81) confirmed that the obtained increase in explicit pro-White bias was driven by more negative evaluations of minorities. Participants' warmth ratings for minorities significantly decreased after IAT score prediction ($M_{Change} = -2.38$, $SE = .63$), $F(1, 237) = 13.94$, $p < .001$, $\eta_p^2 = .056$, but there was no change in the no-prediction condition ($M_{Change} = -0.53$, $SE = .65$), $F(1, 237) = .66$, $p = .419$, $\eta_p^2 = .003$. Ratings of Whites were unaffected by the prediction manipulation, both Fs < 0.80, ps $> .38$ (individual ratings are presented in online supplemental materials Section B).

Effects of feedback. The lack of an effect of IAT Feedback on the magnitude of explicit pro-White bias is particularly interesting in this study because participants in the feedback condition were told that they were biased—and in many cases more biased than they predicted. However, feedback may show effects only for participants whose feedback indicated greater levels of bias than they assume they hold. To test whether extremity of feedback

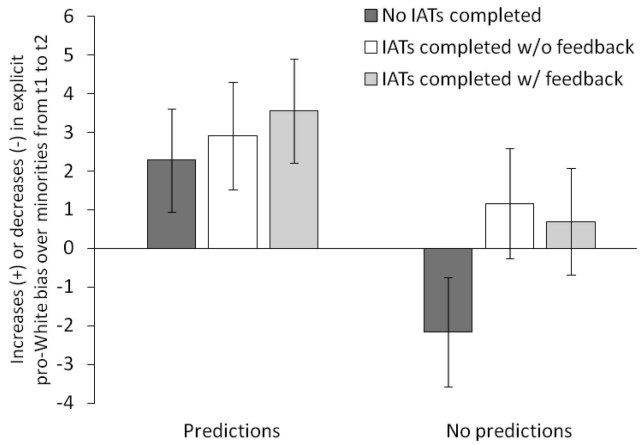

Figure 4. Changes in explicit pro-White bias over Asians, Blacks, and Latinos from Time 1 to Time 2 as a function of IAT score prediction and IAT completion, Study 3. Error bars represent standard errors of predicted values calculated from a 2 (Time) × 2 (IAT Score Prediction) × 3 (IAT Completion) × 3 (Target Group) ANOVA.

influenced explicit preferences among the 83 participants in the feedback condition, we regressed changes in explicit pro-White bias (average of Time 2 preference ratings minus average Time 1 preference ratings) onto a continuous measure of the verbal feedback. Although the relation was in the expected direction with more negative feedback indicating greater increases in explicit bias, this relation did not reach statistical significance, $b = 1.52$, $SE = 1.16$, $t(79) = 1.31$, $p = .194$.

We also tested whether the discrepancy between participants' predictions and their individual feedback was related to changes in explicit bias, equating the seven choices on the prediction scales with the six feedback options (e.g., "slightly more positive towards White" was treated as equivalent to "slight automatic preference for White;" the choice "same reaction" on the prediction scales was treated as equivalent to the feedback options "little to no automatic preference" for either Black or White). For the 42 participants who predicted their scores and received feedback, the discrepancy between predictions and feedback was not significantly related to changes in explicit bias, and the slope of the relation was in the opposite direction, $b = -0.37$, $SE = 2.10$, $t(40) = -0.18$, $p = .86$. Hence, there was no indication that the feedback participants received had any effect on the magnitude of their explicit biases.

Discussion

Study 3 replicated the main findings of Studies 1 and 2. Participants who predicted their IAT scores reported explicit preferences that were more in line with their implicit preferences after than before the prediction task. They also showed stronger explicit bias in favor of Whites over minorities, again replicating a key finding of Studies 1 and 2. IAT completion without feedback showed similar effects as in Study 2, in that participants who had completed IATs without feedback showed marginally greater alignment between explicit and implicit preferences. However, there was no effect of IAT completion on overall levels of explicit bias in favor of Whites over minorities. IAT completion with feedback did lead to significantly greater alignment of implicit and explicit preferences overall, although the simple effect of alignment in response to feedback in the no-predictions condition was not significant, and not significantly larger than the effects of IAT completion without feedback. Feedback did not lead to greater reported levels of bias either, and the content of the IAT feedback was unrelated to mean-levels of explicit preferences. Together, these findings provide further support for the idea that acknowledgment of bias can be increased by directing people's attention to their spontaneous affective reactions (e.g., by asking them to predict IAT scores). Giving people an opportunity to observe behavioral effects of their attitudes via IAT completion (without feedback) may have similar effects, but such effects seem to be smaller in size, less consistent across criterion measures (i.e., marginally increased alignment between explicit and implicit biases vs. no effect on overall levels of explicit biases) and less reliable across studies (i.e., no effect in Study 1, marginal effects in Studies 2 and 3). Effects of IAT feedback were similarly inconsistent (significant effect on alignment compared with control overall, but no effect on size of bias), which is in line with earlier findings showing that many participants may respond defensively to such feedback (see Howell et al., 2015; Howell & Ratliff, 2017).

Although the current findings are consistent with the idea that IAT score prediction increases acknowledgment of bias, the obtained increase in explicit biases and greater alignment between explicit and implicit preferences may reflect reduced concerns about openly expressing one's thoughts and feelings toward minorities. In this case, participants may not necessarily think of their openly expressed judgments as being biased—counter to the proposed interpretation in terms of increased acknowledgment of personal bias. To provide more compelling evidence for the proposed interpretation, Studies 4–6 measured acknowledgment of bias more directly by asking participants to rate the extent to which they harbor racial biases.

Study 4

Study 4 had three aims. The first aim was to test whether the obtained effects of IAT score prediction indeed reflect increased acknowledgment of bias. Toward this end, participants in Study 4 were asked to directly rate the extent to which they harbor racial biases after the manipulation of IAT score prediction. The second aim was to test whether the findings of Studies 1–3 replicate in a more economical online study using a single standard IAT (instead of five shortened IATs), providing a more viable design for potential interventions. Toward this end, participants predicted their performance on a Black–White IAT (or not) and then completed a standard Black–White IAT (or not) before they rated their level of racial bias. The third aim was to investigate whether the effects of IAT score prediction depend on nonprejudicial goals. This aim was based on the assumption that attention to one's spontaneous affective reactions toward minority groups may increase acknowledgment of bias only for participants who endorse nonprejudicial goals. For participants who do not endorse nonprejudicial goals, recognizing negative affective reactions toward minorities may not conflict with personal standards, which should undermine the predicted increase in acknowledgment of bias (Monteith & Mark, 2005). To test this hypothesis, all participants completed a scale measuring nonprejudicial goals at the beginning of the study (Gawronski, Peters, Brochu, & Strack, 2008).

Method

Participants and design. The study included a 2 (IAT Score Prediction: prediction vs. no prediction) × 2 (IAT Completion: completion vs. no completion) between-subjects design, using individual differences in nonprejudicial goals as a continuous moderator. Not having an empirical basis for potential effect sizes in the modified design of Study 4, we aimed to recruit 400 participants, which provides a power of 80% to detect a significant effect of $f = 0.14$. Participants were recruited via TurkPrime and received US $1 for completing the study. Out of 430 participants who began participating in the study, 401 completed all components. Of these participants, 19 failed at least one of two attention check items embedded in the nonprejudicial goals and acknowledgment of bias scales (see Oppenheimer, Meyvis, & Davidenko,

2009)[14], and an additional 24 responded faster than 300 ms on more than 10% of the trials in the IAT (see Greenwald et al., 2003). These participants were excluded from analyses. Of the remaining 358 participants (53.6% female, median age = 34 years, age range = 19–81 years), 77.1% identified as White, 8.9% as Black, 4.2% as Latino, 4.7% as Asian, and the remaining 5.0% as another ethnicity or a combination of several ethnicities.

Nonprejudicial goals. To investigate whether acknowledgment of bias in reaction to predicting IAT scores depends on individual differences in nonprejudicial goals, participants completed the 10 items of Gawronski, Peters, Brochu, and Strack's (2008) nonprejudicial goals scale on 7-point response options ranging from 1 (*strongly disagree*) to 7 (*strongly agree*; Cronbach's alpha = .89, sample item: *Negative evaluations of disadvantaged minority groups are wrong*).

Prediction task. The IAT score prediction task was similar to the one in Studies 1–3. Participants received a short introduction to the concept of *implicit attitudes*, describing them as *a different kind of attitudes*. The text further encouraged participants to *think of them as your spontaneous reactions toward different groups, people, or other targets. Those may be different from the explicit attitudes you would report when you have had time to think about them*. Participants were further told that we were interested in whether they knew their implicit attitudes, and asked them to make a prediction for their implicit attitudes toward cats and dogs, before continuing to a prediction of attitudes toward social groups. The BLACK–WHITE IAT score prediction task included the 20 pictures of Black and White targets used in the IAT with an explanatory text detailing that participants would later complete the IAT. Similar to Studies 1–3, the prediction item read *I predict that an IAT comparing my reactions to BLACK versus WHITE will show that my implicit attitude is . . .* and a 7-point scale ranging from 1 (*a lot more positive toward BLACK*) to 7 (*a lot more positive toward WHITE*). Participants in the control condition completed two of the five consumer preference questions of Studies 1–3 using similar 7-point scales. These questions did not mention IATs or social groups.

Black–White IAT. We used the IATgen tool (https://iatgen.wordpress.com/, Carpenter et al., 2017) to build and implement a seven-block IAT (Greenwald et al., 1998) in Qualtrics, using the pictures of Black and White targets that were also used in Studies 1–3. In Block 1, participants completed 20 trials practicing the categorization of pictures of 10 White and 10 Black people (five male and five female each) using the *I* and *E* keys on their computer keyboards. In Block 2, they completed 20 trials practicing the categorization of positive and negative words using the same keys. Blocks 3 and 4 consisted of 20 and then 40 trials in which participants responded to pictures and words using either a prejudice-compatible or a prejudice-incompatible key mapping. In Block 5, participants practiced the categorization of positive and negative words on 40 trials using a reversed key mapping compared with Block 1. In Blocks 6 and 7, participants completed 20 and then 40 trials categorizing pictures and words with key mappings that were reversed in comparison with Blocks 3 and 4 (prejudice-incompatible or prejudice-compatible). The order of the combined blocks (compatible first vs. incompatible first) and key mappings for Black and White participants (White-left and Black-right vs. Black-left and White-right) were counterbalanced across participants. When participants made an error, they were asked to correct their responses by pressing the other button. Response times were recorded from stimulus onset until participants provided the correct response (see Greenwald et al., 2003). Following standard conventions (Greenwald et al., 2003), IAT *D* scores were computed by calculating the differences between RTs on one incompatible and one compatible block for each participant and dividing them by the pooled standard deviations of those two blocks. This was done once for Blocks 3 and 6, and once for Blocks 4 and 7. The final *D* score reflects the average of those two scores. The IAT showed satisfactory reliability (Cronbach's alpha = .79).[15]

Acknowledgment of bias. Participants completed an eight-item acknowledgment of bias (AoB) scale created for the purpose of the current studies. The scale was designed to capture participants' self-assessment of their automatic racial biases, using the automaticity features of unintentionality, efficiency, and uncontrollability (see Bargh, 1994). Because Hahn et al.'s (2014) findings indicate that people can predict the patterns of their IAT scores with a high degree of accuracy, we did not include the automaticity feature of unawareness (e.g., "unconscious") in the scale. The items of the AoB scale are presented in the Appendix.[16] Responses were measured on 7-point rating scales ranging from 1 (*strongly disagree*) to 7 (*strongly agree*) using individually randomized orders for each participant. Negatively framed items were reverse coded, such that higher numbers on aggregate scores of the AoB scale reflect greater acknowledgment of bias (Cronbach's alpha = .95).

Procedure. After providing informed consent, participants completed the nonprejudicial goals scale. Next, half of the participants predicted their IAT scores, whereas the other half completed the filler task. Next, half of the participants completed the Black–White IAT before completing the AoB scale, whereas the other half completed the AoB scale and then the Black–White IAT. The study concluded with the measurement of demographic information.

Results

Prediction accuracy. In Hahn et al.'s (2014) research, accuracy was relatively high for the prediction of individual patterns of group preferences (e.g., stronger bias against one target group compared with another target group), but lower for the prediction of a given group preference compared with other participants (e.g., stronger bias against a given target group compared with other participants). Whereas the former finding was reflected in rela-

[14] The wording of the two attention check items was: *It is important for us to know whether you read these questions carefully. If you are reading this question carefully, please press 2*, and *This is an attention check. If you are reading this statement, click option 6*.

[15] We calculated the reliability using the shinyapp tool offered by the IATgen webpage at https://applibs.shinyapps.io/iatui2/. It calculates a split-half reliability, corrected by the Spearman-Brown formula, equivalent to a Cronbach's α value based on two individual items.

[16] The items were chosen on the basis of a pilot study in which 119 participants rated their agreement with 13 potential test items. All 13 items loaded >.65 on the same first principal component in a principal component analysis, a scree plot also suggested one factor, and this first factor explained 71% of the variance. For the sake of brevity, we selected eight items that provided a good mix of positively and negatively framed items and could be applied flexibly without reference to particular racial groups.

tively large within-subjects correlations between predicted and actual IAT scores (i.e., high accuracy in predicting one's patterns of IAT scores in a set of five IATs), the latter findings was reflected in medium-size between-subjects correlations between predicted and actual IAT scores (i.e., moderate accuracy in how well participants' predictions reflected how they would score on a given IAT compared with the IAT scores of other participants in the sample). Because the current study included only one IAT, prediction accuracy can be assessed only in terms of between-subjects correlations, but not in terms of within-subjects correlations. Controlling for the counterbalancing of IAT block order and key mapping, the standardized relationship between predicted and actual IAT scores was $r = .28$ (95% CI [.14, .43]), comparable with the medium-size between-subjects correlations reported by Hahn et al. (2014).

IAT scores. A 2 (IAT Score Prediction) × 2 (IAT Completion) ANOVA on IAT scores did not reveal any significant main or interaction effects, all Fs < 0.50, all ps $> .48$, suggesting that IAT scores were unaffected by the prediction task and the order in which participants completed the IAT and the AoB scale.

Acknowledgment of bias. Submitted to a 2 (IAT Score Prediction) × 2 (IAT Completion) ANOVA, AoB scores showed a significant main effect of IAT score prediction, $F(1, 354) = 14.44$, $p < .001$, $\eta_p^2 = .039$. Consistent with the results of Studies 1–3, participants who predicted their IAT scores showed greater acknowledgment of bias than participants who did not predict their IAT scores (see Figure 5). There was no significant main effect of IAT completion, $F(1, 354) = 1.79$, $p = .18$, $\eta_p^2 = .005$, and no significant interaction of IAT score prediction and IAT completion, $F(1, 354) = 0.00$, $p = .99$, $\eta_p^2 = .000$.

Nonprejudicial goals. To investigate whether the effect of IAT score prediction depends on nonprejudicial goals, we regressed AoB scores onto a contrast-coded predictor of the IAT score prediction manipulation (−1 = no prediction, 1 = prediction), a contrast-coded predictor of the IAT completion manipulation (−1 = no IAT, 1 = IAT), z-standardized scores of the nonprejudicial goals scale, as well as all interactions between these three predictors. In addition to replicating the main effect of IAT score prediction, $b = .29$, $SE = .074$, $t(350) = 3.92$, $p < .001$, $\eta_p^2 = .042$, the analysis revealed a significant main effect of nonprejudicial goals, $b = -.21$, $SE = .074$, $t(350) = -2.90$, $p = .004$, $\eta_p^2 = .023$, and a significant interaction of the two predictors, $b = .16$, $SE = .074$, $t(350) = 2.22$, $p = .027$, $\eta_p^2 = .014$. As shown in Figure 6, only participants who scored high on nonprejudicial goals showed greater acknowledgment of bias after predicting IAT scores, $F(1, 350) = 18.79$, $p < .001$, $\eta_p^2 = .051$. There was no effect of the prediction task for participants who scored low on nonprejudicial goals, $F(1, 350) = 1.46$, $p = .228$, $\eta_p^2 = .004$.

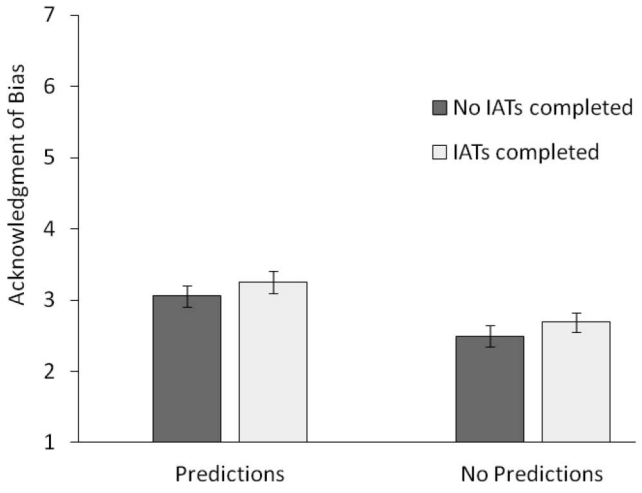

Figure 5. Acknowledgment of bias as a function of IAT score prediction and IAT completion, Study 4. Errors bars depict standard errors of estimated marginal means from a 2 (IAT Score Prediction vs. No Prediction) × 2 (IAT Completed vs. IATs Not Completed) ANOVA.

Discussion

Different from the outcome measures in Studies 1–3, participants in Study 4 were asked to directly rate the extent to which they harbor racial biases after the manipulation of IAT score prediction. Results confirmed our proposed interpretation that increased explicit biases and greater alignment between explicit and implicit preferences reflect increased acknowledgment of bias: Participants who predicted their IAT scores later rated themselves as being more biased compared with participants who did not predict their IAT scores. Moreover, effects of IAT score prediction depended on nonprejudicial goals, in that the prediction task increased acknowledgment of bias for participants with strong nonprejudicial goals, but not for participants with weak nonprejudicial goals. IAT completion did not have any significant effects on self-reported acknowledgment of bias. Another noteworthy aspect of Study 4 is that it replicated the findings of Studies 1–3 in a more economical online design using a single standard IAT (instead of five shortened IATs). This simplified design not only supports the reliability of the obtained effects of IAT score prediction; it also provides a more viable design for potential interventions.

Study 5

Study 5 had three goals. One was to replicate the findings of Study 4 without measuring nonprejudicial goals at the beginning of the study. Our reasoning was that completion of the nonprejudicial goals scale could potentially increase the salience of those goals and thus distort effects of IAT score prediction and IAT completion. Hence, participants in Study 5 went through the same procedure as participants in Study 4 the only difference being that they did not complete the nonprejudicial goals scale. A second goal was to examine the effect of IAT feedback on acknowledgment of bias. Although Study 3 suggests that IAT feedback does not increase overall levels of explicit biases and the alignment between explicit and implicit preferences, we aimed to confirm whether this finding generalizes to a direct measure of acknowledgment of bias. Finally, a third goal was to address a minor programming error in Study 4. Different from the typical IAT procedure in which the key assignment for the target groups (Black vs. White) is reversed in the second set of combined blocks (Greenwald et al., 1998), the IAT in Study 4 was programmed such that the key assignment of the evaluative attributes (good vs. bad) was reversed. This programming error was fixed in Study 5.

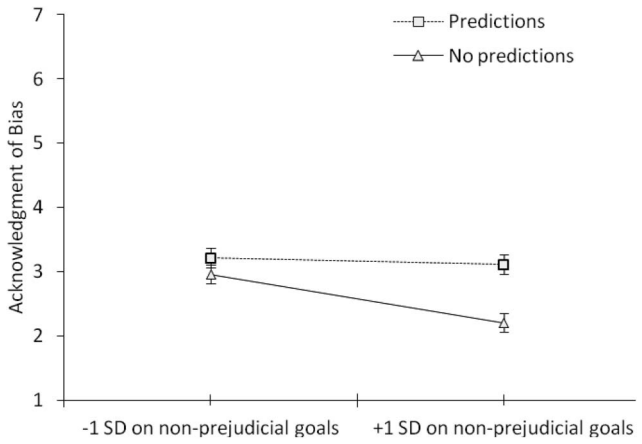

Figure 6. Acknowledgment of bias as a function of nonprejudicial goals and IAT score prediction, Study 4. Errors bars depict standard errors of estimated marginal means from a regression analysis predicting acknowledgment of bias from IAT score prediction, IAT completion, a z-standardized score of Gawronski et al.'s (2008) nonprejudicial goals scale, and all three possible interactions of those predictors.

Method

Participants and design. The study included a 2 (IAT Score Prediction: prediction vs. no prediction) × 3 (IAT Completion: not completed vs. completed without feedback vs. completed with feedback) between-subjects design. Based on the effect sizes in Study 4, a sample size of $N = 260$ provides a power of 90% to replicate the effect of IAT score prediction on acknowledgment of bias. In Study 5, we aimed for a slightly larger sample (a) to compensate for potential overestimations of the obtained effect sizes, (b) to have enough participants per cell with the addition of the two feedback conditions, (c) to have sufficient power to obtain potentially smaller interaction effects, and (d) to have enough participants after excluding fast responders on the IAT. Based on these considerations, we aimed to recruit 480 participants (80 per cell) via TurkPrime in compensation for US $1. Out of 527 participants who initially began the study, 484 completed all components. Data from 35 participants were excluded from analyses. For two participants no IAT data were stored; eight participants failed the attention check item embedded in the AoB scale, and 25 participants responded in less than 300 ms to more than 10% of the trials on the IAT. Of the remaining 449 participants (51.2% female, median age = 33 years, age range = 18–76 years,), 71.9% identified as White, 7.3% as Black, 6.2% as Latino, 7.3% as Asian, and the remaining 7.1% as another ethnicity or as several ethnicities.

Measures. The measures and materials were identical to those in Study 4, except for some minor changes in the IAT. First, we corrected the programming error in Study 4, such that the key assignment for the two target groups (rather than the evaluative attributes) was reversed in the second set of combined blocks (Greenwald et al., 1998). Second, we altered the JavaScript codes from IATgen (Carpenter et al., 2017) to automatically calculate an IAT *D* score (Greenwald et al., 2003). For one third of the participants, the calculated *D* scores were translated into a feedback statement using the same cut-offs and wording as in Study 3 and on the IAT web page. Reliability was satisfactory for the IAT (Cronbach's alpha = .67)[17] as well as the AoB scale (Cronbach's alpha = .94).

Procedure. After participants provided informed consent, roughly half completed the IAT score prediction task and the other half completed the filler task. Next, roughly one third of participants completed the IAT without feedback and one third with feedback. Both of these groups completed the AoB after completing the IAT. The remaining third completed the AoB scale and then the IAT. The study concluded with questions about demographic information.

Results

Prediction accuracy. Controlling for IAT block order and key assignment, the standardized between-subjects relationship between predicted and actual IAT scores was $r = .33$ (95% CI [.20, .45]), slightly higher than in Study 2, but again comparable with the between-subjects correlations reported by Hahn et al. (2014).

IAT scores. A 2 (IAT Score Prediction) × 3 (IAT Completion) ANOVA on IAT scores did not reveal any significant main or interaction effects of the experimental conditions, all Fs < 1.30, all ps $> .29$, suggesting that IAT scores were unaffected by the two manipulations.

Acknowledgment of bias. A 2 (IAT Score Prediction) × 3 (IAT Completion) ANOVA on AoB scores revealed a significant main effect of IAT score prediction, $F(1, 443) = 8.08$, $p = .005$, $\eta_p^2 = .018$. Replicating the findings of Study 4, acknowledgment of bias was greater when participants predicted their IAT scores than when they did not predict their IAT scores (see Figure 7). This main effect was qualified by a significant two-way interaction between IAT score prediction and IAT completion, $F(2, 443) = 3.12$, $p = .045$, $\eta_p^2 = .014$. Inspection of this interaction revealed that IAT completion influenced AoB scores in the no-prediction condition, $F(2, 443) = 3.97$, $p = .018$, $\eta_p^2 = .018$, but not in the prediction condition, $F(2, 443) = .30$, $p = .74$, $\eta_p^2 = .001$ (see Figure 7). More specific contrasts further showed that, among participants who had not predicted their IAT scores, those who completed the IAT with feedback showed significantly higher AoB scores than those who had not completed the IAT, $t(443) = 2.81$, $p = .005$, $\eta_p^2 = .017$. Participants in the no-prediction condition who had completed the IAT without feedback showed AoB scores in-between the two groups, differing neither from participants in the no-completion condition, $t(443) = 1.63$, $p = .104$, $\eta_p^2 = .006$, nor from participants in the IAT with feedback condition, $t(443) = 1.23$, $p = .218$, $\eta_p^2 = .003$. The interaction was also evident in that the effect of prediction on acknowledgment was significant in the no-IAT condition, $t(443) = 3.59$, $p < .001$, $\eta_p^2 = .028$, but not in the IAT plus feedback condition, $t(443) = .09$, $p = .930$, $\eta_p^2 = .000$. An interaction of a specific contrast comparing these two effects of the prediction manipulation was significant, $t(443) = -2.45$, $p = .015$, $\eta_p^2 = .013$. The effect of prediction was not significant in the IAT without feedback condition, $t(443) = 1.28$, $p = .203$, $\eta_p^2 = .004$, although interaction

[17] Reliability was again calculated as a Cronbach's α value of the two separate *D* scores (i.e., their correlation corrected by the Spearman-Brown formula).

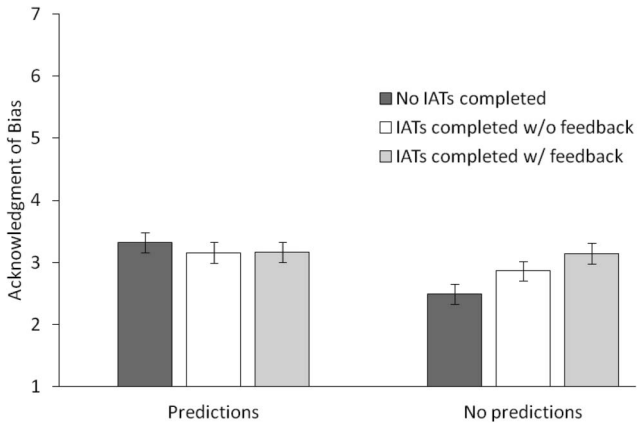

Figure 7. Acknowledgment of bias as a function of IAT score predictions and IAT completion, Study 5. Errors bars depict standard errors of estimated marginal means from a 2 (IAT Score Predictions vs. No Predictions) × 2 (IAT Completed With Feedback vs. IATs Completed Without Feedback vs. No IATs Completed) ANOVA.

effects with more specific contrasts revealed that it was neither significantly smaller than the effect in the no-IAT condition, $t(443) = -1.64$, $p = .102$, $\eta_p^2 = .006$, nor significantly bigger than in the IAT with feedback condition, $t(443) = -.83$, $p = .407$, $\eta_p^2 = .002$.

Effects of feedback. To investigate whether the extremity of feedback influenced acknowledgment of bias among the 146 participants in the feedback condition, we created a continuous measure of the verbal feedback in absolute terms (from 0 = *little to no automatic preference* to 3 = *strong automatic preference*). We then regressed AoB scores onto a z-standardized measure of this score, a contrast comparing the prediction (coded 1) with the no-prediction condition (coded −1), and their interaction. Results revealed no significant relationship between feedback and AoB scores, and the size of the slope suggested, if anything, a negative relationship, $b = -.08$, $SE = .112$, $t(142) = -.71$, $p = .479$. The lack of a significant interaction revealed that this was true for both the predictions and the no-predictions conditions, $b = .03$, $SE = .112$, $t(142) = .22$, $p = .826$.

Following the analyses in Study 3, we also created a discrepancy score for the 73 participants who predicted their scores and received feedback between their absolute feedback scores and their absolute prediction scores. Surprisingly, the discrepancy between predictions and feedback was significantly *negatively* related to acknowledgment of bias, $b = -.27$, $SE = .114$, $t(71) = -2.33$, $p = .023$, standardized relationship: $r = -.27$. Hence, the more negative participants' feedback was compared with their predictions, the less they agreed with items describing them as biased, reminiscent of Howell and colleagues' findings on defensive responding to IAT feedback (Howell et al., 2015; Howell & Ratliff, 2017). Thus, although IAT completion with feedback led to an overall increase in acknowledgment of bias when participants did not predict IAT scores, acknowledgment of bias was unrelated to the individual feedback participants received, and it was related to reduced acknowledgment when it exceeded participants' expectations in the prediction condition.

Discussion

In Study 5, participants went through the same procedure as participants in Study 4, one important difference being that they did not complete a measure of nonprejudicial goals prior to the manipulation of IAT score prediction. Despite this procedural difference, IAT score prediction increased acknowledgment of bias, replicating the main finding of Study 4. IAT completion without feedback had no significant effect on acknowledgment of bias, again replicating the results of Study 4. Yet, different from the findings in Study 3, where IAT completion with feedback had no effect on the overall size of explicit preferences, IAT completion with feedback increased acknowledgment of bias in the current study. However, this increase was unrelated to the individual feedback participants received on their IAT performance, and it was related to reduced acknowledgment to the degree that it contradicted participants' expectations for how much bias they would show. This result suggests that those who would have the most to learn from receiving IAT feedback would be least likely to accept it.

Study 6

Studies 1–5 suggest that predicting one's IAT scores can increase acknowledgment of bias. This was reflected in (a) increased alignment between explicit and implicit preferences, (b) greater levels of explicit biases, and (c) enhanced self-reports of harboring racial biases. These findings are consistent with the idea that IAT score prediction enhances attention to one's spontaneous affective reactions toward minority members, which increases acknowledgment of bias by making those reactions more salient and counteracting their dismissal. However, it is also compatible with an interpretation suggesting that anticipating the completion of a psychological test that will uncover one's biases increases people's willingness to admit these biases. In the current studies, participants who were asked to predict their IAT scores were also told that they would later complete the IATs for which they were asked to predict their scores, and this announcement may be a necessary ingredient for the effectiveness of the procedure.

The main goal of Study 6 was to address this confound. Toward this end, we asked one group of participants to predict their IAT scores as in all previous studies (prediction condition). Another group of participants was asked to attend to their spontaneous affective reactions toward minority groups without asking them to make any predictions, and without reference to any of the terminology used in discussions surrounding implicit bias (e.g., "unconscious" or "implicit," attention condition). We then compared acknowledgment of bias among participants in the two conditions with each other as well as to a third group of participants who was neither asked to predict IAT scores, nor asked to attend to spontaneous affective reactions (control condition). To the extent that the findings in Studies 1–5 can be explained by enhanced attention to spontaneous affective reactions during the prediction task alone (and not by test announcement), acknowledgment of bias should be greater in both the prediction and the attention conditions compared with the control condition. Yet, if anticipation of a psychological test that will reveal one's personal biases is a necessary component of the acknowledgment of bias effects observed in Studies 1–5, acknowledgment of bias should be higher in the

predictions as opposed to both the attention and the control conditions.

Method

Participants and design. The study used a one-factorial design with three between-subjects conditions: (a) IAT score prediction, (b) attention to spontaneous affective reactions, and (c) control condition. Power analyses suggested a sample of 260 participants to provide a probability of 90% to replicate the significant effect of IAT score prediction in Study 4, and a sample of 365 to replicate the simple effect of IAT score prediction in the no-completion condition of Study 5. Based on these estimations and anticipated exclusions, we aimed to recruit 390 participants (130 per condition) on TurkPrime for a compensation of US $0.50. Out of 402 participants who started completing the study, 392 completed all measures. Of these participants, nine failed the attention check item and were therefore excluded from analyses. Of the remaining 383 participants (50.9% female, median age = 33 years, age range 18–74 years), 79.9% identified as White, 5.7% as Black, 4.2% as Latino, 4.2% as Asian, and the remaining 6% as another ethnicity or several ethnic backgrounds.

Materials and procedure. After providing informed consent, participants were randomly assigned to one of the three experimental conditions. Participants in the current study did not complete any IATs. Materials in the IAT score prediction condition and the control condition were similar to those of Studies 4 and 5. Participants in the prediction condition completed the IAT score prediction task and participants in the control condition completed the consumer preference task. Participants in the attention condition received the following instructions:

> Psychologists have long been interested in people's spontaneous reactions toward different people. That is, 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. In this study, we are interested in your first reactions. In a minute, you will see pictures of people and we want to know what your first reaction is, independent of what you would say once you have had time to reflect about your opinion. Please be honest and take your time to observe how you feel in the first second when you look at the pictures.

As in the prediction condition, participants then provided a "test indication" about their reactions to cats and dogs, introduced as follows:

> Before you report your spontaneous reactions toward different social groups, we would like you to get used to the scales you will use to report those reactions. Remember that your first reaction could be different from a general opinion you may have. And right now, we are only interested in your first reaction.

The scales included the same pictures of Black and White people formatted in the same way as in the prediction manipulation. A text encouraged participants to look at the pictures and pay attention to their immediate spontaneous reactions. They were then asked to complete the sentence *My spontaneous reaction to BLACK versus WHITE is . . .* by checking one response option on a 7-point rating scale ranging from −3 (*a lot more positive toward BLACK*) to +3 (*a lot more positive toward WHITE*). Hence, in addition to not announcing measurement, the attention manipulation also never mentioned the word "implicit" in any combination to avoid that participants would draw connections to discussions surrounding implicit bias and anticipate a test. After completing either the prediction, the attention, or the filler task, participants completed the AoB scale (Cronbach's alpha = .96), provided demographic information, and were then debriefed. The debriefing included information that they would not be asked to complete IATs in this study (despite the announcement in the prediction condition) and the reason for this deception. Participants were further told that they could do so at the website of project implicit (http://implicit.harvard.edu).

Results

Participants reported similar levels of pro-White bias when they reported their spontaneous affective reactions ($M = .95$, $SD = 1.28$) and when they predicted IAT scores ($M = .73$, $SD = 1.47$). Both of these scores differed significantly from a no-bias score of 0, attention: $t(131) = 8.55$, $p < .001$, $d = .74$ and prediction: $t(121) = 5.50$, $p < .001$, $d = .50$; but they did not significantly differ from each other, $t(252) = -1.30$, $p = .193$, $d = -.16$.

A one-way ANOVA on AoB scores further revealed a significant difference between experimental conditions, $F(2, 380) = 15.02$, $p < .001$, $\eta_p^2 = .073$ (see Figure 8). Simple-effect contrasts revealed that participants who predicted their IAT scores showed greater acknowledgment of bias than participants in the control condition, $F(1, 380) = 15.65$, $p < .001$, $\eta_p^2 = .040$. Similarly, participants who were asked to pay attention to their spontaneous affective reactions showed greater acknowledgment of bias than participants in the control condition, $F(1, 380) = 27.62$, $p < .001$, $\eta_p^2 = .068$. Participants in the attention condition did not differ from participants in the prediction condition, $F(1, 380) = 1.45$, $p = .230$, $\eta_p^2 = .004$. Hence, there was no evidence that attention

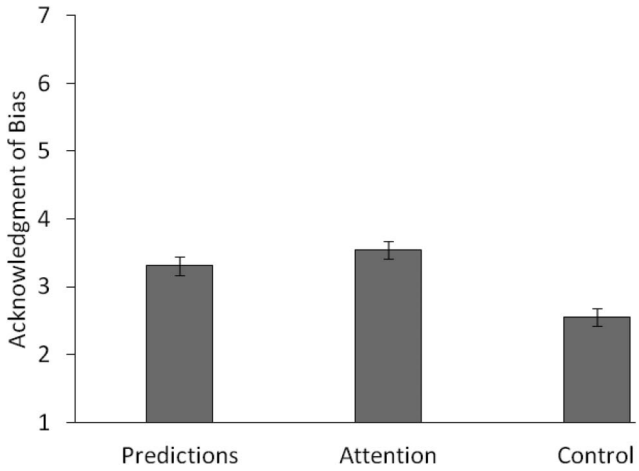

Figure 8. Acknowledgment of bias as a function of IAT score prediction (predictions), attention to affective reactions (attention), or completion of a filler task (control). Study 6. Errors bars depict standard errors of estimated marginal means from a one-way ANOVA testing differences between the three conditions.

without test announcement had weaker effects than predicting IAT scores; the means, if anything, suggested the opposite pattern.

Discussion

The main goal of Study 6 was to provide more compelling evidence for the hypothesis that IAT score prediction increases acknowledgment of bias via enhanced attention to spontaneous affective reactions toward minority group members. Toward this end, we compared the prediction manipulation utilized in Studies 1–5 with a condition where participants were asked to pay attention to their spontaneous reactions without any reference to a psychological test or constructs used in discussions surrounding implicit bias. Consistent with the hypothesis that the findings in Studies 1–5 are driven by enhanced attention to spontaneous affective reactions during the prediction task, and not by anticipation of a test that will reveal one's biases, acknowledgment of bias was greater in both the attention and the prediction conditions compared with the control condition, and even nonsignificantly larger in the attention condition. These findings rule out an interpretation suggesting that anticipation of a psychological bias test is a necessary component for the effects observed in Studies 1–5.

General Discussion

Expanding on conflicting conceptions of implicit bias, the current research investigated the effectiveness of different procedures to increase acknowledgment of harboring biases against minorities, focusing particularly on the effects of predicting one's IAT scores. Studies 1–3 showed that participants who predicted their responses toward various minority groups on future IATs showed increased alignment between implicit and explicit preferences and greater explicit bias against minorities. Expanding on these findings, Studies 4–6 demonstrated that participants who predicted their racial bias on a Black–White IAT later described themselves as harboring greater levels of automatic racial bias. Study 4 further showed that the effect of IAT score prediction depends on the endorsement of nonprejudicial goals, in that the effects of IAT score prediction on acknowledgment of bias were larger for participants with strong nonprejudicial goals, but diminished for participants with weak nonprejudicial goals. Finally, Study 6 showed that instructions to attend to one's spontaneous affective reactions toward minority members increased acknowledgment of bias to the same extent as IAT score prediction.

Completion of IATs with and without feedback had inconsistent effects across studies and criterion measures. Although IAT completion *without* feedback marginally increased alignment between implicit and explicit preferences in Study 2 and 3, this effect was nonsignificant in Study 1. IAT completion without feedback did not increase explicit bias either (and decreased explicit bias in Studies 1 and 2), and showed no significant effects on self-reported acknowledgment of bias.

IAT completion *with* feedback may have increased the alignment between implicit and explicit preferences in Study 3 (there was no simple alignment effect in the feedback condition, but an interaction indicating that alignment was larger than in the control condition), but it did not lead to an increase in reported bias either. It further did lead to an increase in self-reported acknowledgment of bias (Study 5), but this increase was unrelated to the content of the individual feedback, and acknowledgment was negatively related to the content of the feedback to the extent that it exceeded participants' predictions.

Together, these results suggest that acknowledgment of bias can be increased by directing people's attention to their spontaneous affective reactions toward minority groups (e.g., by asking them to predict their scores on future IATs). Effects of IAT completion and IAT feedback on acknowledgment of bias were inconsistent and less reliable.

Theoretical Implications

A popular explanation for differences between implicit and explicit evaluations is that implicit evaluations reflect attitudes people are unable report (e.g., https://implicit.harvard.edu/implicit/education.html). According to this view, people have no introspective access to the attitudes underlying their implicit biases, which makes them unable to report these attitudes on traditional self-report measures (for perspectives challenging this view, see Gawronski et al., 2006; Hahn & Gawronski, 2014: Hahn et al., 2014). Based on this conception, one potential way to increase acknowledgment of bias is to inform people about their implicit biases by providing individual feedback on their IAT scores. This hypothesis is consistent with the finding that IAT feedback increased self-reported acknowledgment of bias (Study 5). However, it is challenged by the findings that (a) the obtained increase in acknowledgment of bias was unrelated to the individual feedback participants received on their IAT performance and negatively related to acknowledgment to the degree that it exceeded participants' performance expectations (Study 5); and (b) IAT feedback had no effect on overall levels of explicit bias (Study 3).

Another popular explanation for differences between implicit and explicit evaluations is that implicit evaluations reflect attitudes people are unwilling report (e.g., https://implicit.harvard.edu/implicit/education.html). This conception is based on the idea that responses on traditional self-report measures are much easier to control than responses on performance-based measures. Hence, honest reporting of one's biases on self-report measures can be undermined by social desirability and other self-presentational concerns (for a critical review, see Gawronski et al., 2007). From this perspective, one potential way to increase acknowledgment of bias is to tell participants that their personal biases will be identified with a performance-based measure that cannot be controlled. Anticipating such a test may increase participants' willingness to admit to their biases (e.g., Nier, 2005) and this increase may occur without actual completion of an implicit bias test or individual feedback on measurement scores. Because the IAT score prediction task in Studies 1–5 informed participants about the completion of future IATs, this hypothesis is consistent with the findings that alignment between implicit and explicit preferences, overall levels of explicit bias, and self-reported acknowledgment of bias were greater when participants were asked to predict their IAT scores. It is also consistent with the finding that IAT completion alone, compared with a control condition that did not announce a test, marginally increased the alignment between explicit and implicit preferences in Studies 2 and 3. However, it is inconsistent with the finding that IAT completion alone did not increase acknowledgment of bias (Studies 4–6) or explicit bias (Studies 1–3). Most importantly, Study 6 suggests that anticipation of a psychological bias

test is not a necessary component of the effects of predicting one's IAT scores. Participants who were simply asked to report their spontaneous affective reactions toward a set of pictures showed the same acknowledgment effect as a group of participants who predicted their IAT scores.

A third possibility is that people can become aware of the unconscious attitudes underlying their implicit biases by observing behavioral effects of their unconscious attitudes (Hofmann et al., 2009; Hofmann & Wilson, 2010). Because participants typically notice the difference in their RTs and errors in the prejudice-congruent and prejudice-incongruent blocks of the IAT (Monteith et al., 2001), mere completion of an IAT may increase acknowledgment of bias to the extent that participants notice the behavioral effects of their attitudes in the task. Different from the hypothesis that people are generally unable to report their implicit biases, the notion of self-perception (Bem, 1972) suggests that merely completing an IAT may increase acknowledgment of bias by observing one's behavior even without feedback about one's measurement scores. This hypothesis is consistent with the finding that IAT completion without feedback marginally increased alignment of implicit and explicit evaluations in Studies 2 and 3. However, in addition to this effect being only marginal in Studies 2 and 3, it was not significant in Study 1. Moreover, IAT completion without feedback did not increase overall explicit bias (Studies 1–3) and it did not increase self-reported acknowledgment of bias (Studies 4–5). Thus, even if IAT completion can contribute to increased acknowledgment of bias via self-perception, such effects seem to be small, unreliable, and inconsistent across outcome measures.

Finally, a fourth conception suggests that implicit evaluations are subjectively experienced as spontaneous affective reactions, and dissociations between implicit and explicit evaluations arise from differences in the extent to which people rely on their spontaneous affective reactions in making an evaluative judgment (Fazio, 2007; Gawronski & Bodenhausen, 2006, 2011). Thus, to the extent that people do not base their overt judgments on spontaneous affective reactions, they may dismiss the significance of these reactions in producing discriminatory behavior. From this perspective, directing people's attention to their spontaneous affective reactions toward minority members may increase acknowledgment of bias by counteracting the dismissal of these reactions. To the extent that prediction of IAT scores enhances such attention, acknowledgment of bias could be increased by asking participants to predict their IAT scores without requiring them to complete an IAT, without feedback on their measurement scores, and without anticipation of actual measurement. This hypothesis is consistent with the finding that IAT score prediction consistently increased acknowledgment of bias on all three criterion measures. It led to increased alignment between implicit and explicit preferences (Studies 1–3), greater levels of explicit bias (Studies 1–3), and increased self-reported acknowledgment of being racially biased (Studies 4–6). Moreover, instructions to attend to one's spontaneous affective reactions toward minority members increased acknowledgment of bias to the same extent as IAT score prediction (Study 6), providing further evidence for the functional equivalence of IAT score prediction and attention to spontaneous affective reactions.

Implications for Interventions

The IAT has become a popular "consciousness raising" tool in educational settings (Casad et al., 2013; Hillard et al., 2013) and the popular media (Dateline NBC, 2007; This American Life, 2015), even reaching into debates on policy applications (Hillaryclinton.com, 2016; Reuters, 2016). Many of these trainings and exercises involve the completion of IATs, feedback on IAT performance, information on the meaning of the implicit bias construct, and extensive discussions about participants' personal biases (e.g., Devine, Forscher, Austin, & Cox, 2012; Forscher, Mitamura, Dix, Cox, & Devine, 2017). By dissociating these aspects experimentally, the current studies showed that the discussion part might be the more effective ingredient in such interventions compared with IAT completion and IAT feedback. In the current studies, directing participants' attention to their spontaneous affective reactions by asking them to predict their scores on future IATs led to increased alignment between implicit and explicit preferences, greater levels of explicit bias, and greater acknowledgment of personal bias. IAT completion and IAT feedback showed inconsistent effects across studies and outcome measures.

Note, however, that the current findings do not provide any information on the temporal persistence of the obtained effects, which remains an important question for future research. Neither do they speak to the question of whether increased acknowledgment of bias is beneficial for intergroup relations. Such benefits are often taken for granted and extant theories persuasively argue that acknowledgment of bias may be an important first step in counteracting prejudice and discrimination (e.g., Monteith & Mark, 2005). However, several additional factors may have to be considered to understand the downstream effects of increased acknowledgment of bias. In the exploratory analyses reported in online supplemental materials Section A, we investigated downstream effects on internal and external motivation to respond without prejudice (Plant & Devine, 1998). Although these analyses suggest that IAT score prediction increases the motivation to respond without prejudice, this effect was eliminated when participants completed IATs in addition to predicting their IAT scores. These exploratory results suggest that the downstream effects of increased acknowledgment of bias on motivation to counteract bias may be more complex and dependent on other factors. Nevertheless, our findings suggest that one particularly effective way of increasing acknowledgment of bias is to direct people's attention to their spontaneous affective reactions toward minority groups. More research is needed to elucidate whether or not such acknowledgment is a beneficial strategy for interventions against discrimination and inequality at the societal level.

Conclusion

Implicit biases are often presented as attitudes people are unable or unwilling to report (e.g., https://implicit.harvard.edu/implicit/education.html). Sometimes, these attitudes are described as unconscious (e.g., Jost, Pelham, & Carvallo, 2002; Phelps et al., 2000), suggesting that conscious awareness of the attitudes underlying implicit biases is impossible (e.g., Devos, 2008; Kassin, Fein, & Markus, 2011; Kihlstrom, 2004; McConnell, Dunn, Austin, & Rawn, 2011). Based on these assumptions, it seems quite remarkable that people can predict their IAT with a high level of accuracy (Hahn et al., 2014) and predicting one's IAT scores increases (a) the alignment between implicit and explicit preferences, (b) overall levels of explicit bias, and (c) self-reported acknowledgment of harboring automatic biases. In light of conceptions that assume that people are unwilling to admit to their biases, it also seems remarkable that instructions to direct one's

attention to one's spontaneous affective reactions, without any announcement of test completion, increased acknowledgment of bias to the same extent as a predicting one's IAT scores.

Despite their inconsistency with the two most common narratives in research on implicit bias, the current findings are consistent with theories suggesting that implicit evaluations are subjectively experienced as spontaneous affective reactions (e.g., Fazio, 2007; Gawronski & Bodenhausen, 2006, 2011). According to these theories, predicting one's IAT scores may increase attention to one's spontaneous affective reaction, which may counteract the dismissal of these reactions. Based on these conclusions, we deem it problematic to present implicit biases as attitudes that people are either unable or unwilling to report. Aside from being difficult to reconcile with the available evidence, such conceptualizations may thwart the path to implementing successful methods of educating the public about effective ways to foster acknowledgment of personal biases. The current findings suggest that a presentation of implicit biases as spontaneous affective reactions may be more accurate, opening the door for the development of more effective bias interventions. As we noted above, whether or not acknowledgment of bias also leads to increased efforts to control one's biases is a question awaiting future research. However, informed debates about bias intervention require that we accept and publicly communicate the fact that differences between implicit and explicit preferences are rooted in factors that have little to do with lack of awareness or dishonest self-reports.

References

Bargh, J. A. (1994). The four horsemen of automaticity: Awareness, intention, efficiency, and control in social cognition. In R. S. Wyer & T. K. Srull (Eds.), *Handbook of social cognition* (pp. 1–40). Hillsdale, NJ: Erlbaum.

Bem, D. J. (1972). Self-perception theory. *Advances in Experimental Social Psychology, 6,* 1–62. http://dx.doi.org/10.1016/S0065-2601(08)60024-6

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience, 14,* 365–376. http://dx.doi.org/10.1038/nrn3475

Carpenter, T., Pogacar, R., Pullig, C., Kouril, M., LaBouff, J., Aguilar, S., . . . Chakroff, A. (2017). *Building and analyzing implicit association tests for online surveys: A tutorial and open-source tool*. Retrieved from osf.io/preprints/psyarxiv/6xdyj

Casad, B. J., Flores, A. J., & Didway, J. D. (2013). Using the implicit association test as an unconsciousness raising tool in psychology. *Teaching of Psychology, 40,* 118–123. http://dx.doi.org/10.1177/0098628312475031

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science, 25,* 7–29. http://dx.doi.org/10.1177/0956797613504966

Czopp, A. M., Monteith, M. J., & Mark, A. Y. (2006). Standing up for a change: Reducing bias through interpersonal confrontation. *Journal of Personality and Social Psychology, 90,* 784–803. http://dx.doi.org/10.1037/0022-3514.90.5.784

Dateline NBC. (2007, April 19). *Dateline NBC: Psychological dispositions in Black & White* [Video file]. Retrieved from https://www.youtube.com/watch?v=n5Q5FQfXZag

Devine, P. G., Forscher, P. S., Austin, A. J., & Cox, W. T. L. (2012). Long-term reduction in implicit race bias: A prejudice habit-breaking intervention. *Journal of Experimental Social Psychology, 48,* 1267–1278. http://dx.doi.org/10.1016/j.jesp.2012.06.003

Devos, T. (2008). Implicit attitudes 101: Theoretical and empirical insights. In W. D. Crano & R. Prislin (Eds.), *Attitudes and attitude change* (pp. 61–84). New York, NY: Psychology Press.

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

Epstein, S., Pacini, R., Denes-Raj, V., & Heier, H. (1996). Individual differences in intuitive-experiential and analytical-rational thinking styles. *Journal of Personality and Social Psychology, 71,* 390–405. http://dx.doi.org/10.1037/0022-3514.71.2.390

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,* 175–191. http://dx.doi.org/10.3758/BF03193146

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

Fazio, R. H., Jackson, J. R., Dunton, B. C., & Williams, C. J. (1995). Variability in automatic activation as an unobtrusive measure of racial attitudes: A bona fide pipeline? *Journal of Personality and Social Psychology, 69,* 1013–1027. http://dx.doi.org/10.1037/0022-3514.69.6.1013

Forscher, P. S., Mitamura, C., Dix, E. L., Cox, W. T. L., & Devine, P. G. (2017). Breaking the prejudice habit: Mechanisms, timecourse, and longevity. *Journal of Experimental Social Psychology, 72,* 133–146. http://dx.doi.org/10.1016/j.jesp.2017.04.009

Frantz, C. M., Cuddy, A. J., Burnett, M., Ray, H., & Hart, A. (2004). A threat in the computer: The race implicit association test as a stereotype threat experience. *Personality and Social Psychology Bulletin, 30,* 1611–1624. http://dx.doi.org/10.1177/0146167204266650

Gawronski, B., & Bodenhausen, G. V. (2006). Associative and propositional processes in evaluation: An integrative review of implicit and explicit attitude change. *Psychological Bulletin, 132,* 692–731. http://dx.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. http://dx.doi.org/10.1016/B978-0-12-385522-0.00002-0

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

Gawronski, B., LeBel, E. P., & Peters, K. R. (2007). What do implicit measures tell us? Scrutinizing the validity of three common assumptions. *Perspectives on Psychological Science, 2,* 181–193. http://dx.doi.org/10.1111/j.1745-6916.2007.00036.x

Gawronski, B., Peters, K. R., Brochu, P. M., & Strack, F. (2008). Understanding the relations between different forms of racial prejudice: A cognitive consistency perspective. *Personality and Social Psychology Bulletin, 34,* 648–665. http://dx.doi.org/10.1177/0146167207313729

German Federal Agency for Civic Education. (2016). *Die soziale Situation in Deutschland, Bevölkerung mit Migrationshintergrund* [The social situation in Germany, population with migrant backgrounds]. Retrieved from http://www.bpb.de/

German Federal Statistical Office. (2015). *Bevölkerung mit Migrationshintergrund—Ergebnisse des Mikrozensus—Fachserie 1 Reihe 2.2–2014* [Population with migrant backgrounds—Results from a microcensus—Content series 1 serial 2.2–2014]. Retrieved from http://www.destatis.de/

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review, 102,* 4–27. http://dx.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. http://dx.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,* 197–216. http://dx.doi.org/10.1037/0022-3514.85.2.197

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

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

Hahn, A., Judd, C. M., & Park, B. (2010). Thinking about group differences: Ideologies and national identities. *Psychological Inquiry, 21,* 120–126. http://dx.doi.org/10.1080/1047840X.2010.483997

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,* 495–516. http://dx.doi.org/10.1007/s11218-013-9219-5

Hillaryclinton.com. (2016, September 28). *For African Americans, implicit bias can be deadly. It's time to address it*. Retrieved from http://www.Hillaryclinton.com

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,* 1369–1385. http://dx.doi.org/10.1177/0146167205275613

Hofmann, W., Gschwendner, T., & Schmitt, M. (2009). The road to the implicit self not taken: Discrepancies between self- and observer-inferences about implicit dispositions from nonverbal behavioral cures. *European Journal of Personality, 23,* 343–366. http://dx.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). New York, NY: Guilford Press.

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 & Personality Science, 6,* 373–381. http://dx.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,* 125–145. http://dx.doi.org/10.1111/bjso.12168

Jost, J. T., Pelham, B. W., & Carvallo, M. R. (2002). Non-conscious forms of system justification: Implicit and behavioral preferences for higher status groups. *Journal of Experimental Social Psychology, 38,* 586–602. http://dx.doi.org/10.1016/S0022-1031(02)00505-X

Kassin, S., Fein, S., & Markus, H. R. (2011). *Social psychology* (8th ed.). Belmont, CA: Wadsworth Cengage Learning.

Kihlstrom, J. F. (2004). Implicit methods in social psychology. In C. Sansone, C. C. Morf, & A. T. Panter (Eds.), *The Sage handbook of methods in social psychology* (pp. 195–212). Thousand Oaks, CA: Sage.

McConahay, J. B. (1986). Modern racism, ambivalence, and the modern racism scale. In J. D. Dovidio & S. L. Gaertner (Eds.), *Prejudice, discrimination, and racism* (pp. 91–125). San Diego, CA: Academic Press.

McConnell, A. R., Dunn, E. W., Austin, S. N., & Rawn, C. D. (2011). Blind spots in the search for happiness: Implicit attitudes and nonverbal leakage predict affective forecasting errors. *Journal of Experimental Social Psychology, 47,* 628–634. http://dx.doi.org/10.1016/j.jesp.2010.12.018

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

Monteith, M. J., Devine, P. G., & Zuwerink, J. R. (1993). Self-directed versus other-directed affect as a consequence of prejudice-related discrepancies. *Journal of Personality and Social Psychology, 64,* 198–210. http://dx.doi.org/10.1037/0022-3514.64.2.198

Monteith, M. J., & Mark, A. Y. (2005). Changing one's prejudiced ways: Awareness, affect, and self-regulation. *European Review of Social Psychology, 16,* 113–154. http://dx.doi.org/10.1080/10463280500229882

Monteith, M. J., Voils, C. I., & Ashburn-Nardo, L. (2001). Taking a look underground: Detecting, interpreting, and reacting to implicit racial biases. *Social Cognition, 19,* 395–417. http://dx.doi.org/10.1521/soco.19.4.395.20759

Nier, J. A. (2005). How dissociated are implicit and explicit racial attitudes? A bogus pipeline approach. *Group Processes & Intergroup Relations, 8,* 39–52. http://dx.doi.org/10.1177/1368430205048615

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

Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology, 45,* 867–872. http://dx.doi.org/10.1016/j.jesp.2009.03.009

Payne, B. K., Cheng, C. M., Govorun, O., & Stewart, B. D. (2005). An inkblot for attitudes: Affect misattribution as implicit measurement. *Journal of Personality and Social Psychology, 89,* 277–293. http://dx.doi.org/10.1037/0022-3514.89.3.277

Petty, R. E., Fazio, R. H., & Briñol, P. (Eds.). (2009). The new implicit measures: An overview. *Attitudes: Insights from the new implicit measures* (pp. 3–18). New York, NY: Psychology Press.

Phelps, E. A., O'Connor, K. J., Cunningham, W. A., Funayama, E. S., Gatenby, J. C., Gore, J. C., & Banaji, M. R. (2000). Performance on indirect measures of race evaluation predicts amygdala activation. *Journal of Cognitive Neuroscience, 12,* 729–738. http://dx.doi.org/10.1162/089892900562552

Plant, E. A., & Devine, P. G. (1998). Internal and external motivation to respond without prejudice. *Journal of Personality and Social Psychology, 75,* 811–832. http://dx.doi.org/10.1037/0022-3514.75.3.811

Reuters. (2016, June 27). *J. Edwards: Justice department mandates 'implicit bias' training for agents, lawyers*. Retrieved from http://www.reuters.com/article/us-usa-justice-bias-exclusive-idUSKCN0ZD251

This American Life. (2015, February 13). *548: Cops see it differently, part two. This American Life* [Audio podcast]. Retrieved from http://www.thisamericanlife.org/radio-archives/episode/548/cops-see-it-differently-part-two

Uhlmann, E. L., & Nosek, B. A. (2012). My culture made me do it. *Social Psychology, 43,* 108–112. http://dx.doi.org/10.1027/1864-9335/a000089

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology, 54,* 1063–1070. http://dx.doi.org/10.1037/0022-3514.54.6.1063

(*Appendix follows*)

Appendix

Acknowledgment of Bias Scale Used in Studies 4–6 and Factor Loading on First Principal Component in Principal Component Analysis (PCA)

Scale items	Factor loadings on first principal component in PCA		
	Study 4	Study 5	Study 6
1. I have negative biases against other racial groups.	.83	.78	.84
2. Whether I want it or not, my spontaneous reactions towards people are racially biased.	.91	.90	.94
3. I have an unintentional racial bias in my first reactions towards strangers.	.86	.79	.82
4. I show no racial bias in my reactions towards other people. (rev.)	.80	.86	.90
5. My automatic reactions towards other people are racially biased.	.88	.89	.92
6. When I observe my own spontaneous reactions when meeting strangers, I see no racial bias. (rev.)	.84	.84	.85
7. My immediate feelings when I encounter new people often show racial biases.	.83	.85	.89
8. My first reactions to other people are not influenced by their racial background. (rev)	.84	.86	.85
% of variance explained by first component in PCA	72.1	71.8	76.6
Cronbach's α	.95	.94	.96

Received May 19, 2017
Revision received May 16, 2018
Accepted July 7, 2018 ■