AMERICAN PSYCHOLOGICAL ASSOCIATION

Journal of Experimental Psychology: General

ISSN: 0096-3445

2023, Vol. 152, No. 12, 3311–3343
https://doi.org/10.1037/xge0001464

Awareness of Implicit Attitudes: Large-Scale Investigations of Mechanism and Scope

Adam Morris[1] and Benedek Kurdi[2]
[1] Department of Psychology, Princeton University
[2] Department of Psychology, University of Illinois Urbana-Champaign

People can predict their scores on the Implicit Association Test with remarkable accuracy, challenging the traditional notion that implicit attitudes are inaccessible to introspection and suggesting that people might be aware of these attitudes. Yet, major open questions about the mechanism and scope of these predictions remain, making their implications unclear. Notably, people may be inferring their attitudes from externally observable cues (e.g., in the simplest case, their demographic information) rather than accessing them directly. This problem is exacerbated by the fact that, in past work, predictions have been obtained only for a small set of targets, attitudes toward which are demonstrably possible to infer. Here, in an adversarial collaboration with eight preregistered studies ($N = 8{,}011$), we interrogate implicit attitude awareness using more stringent tests. We demonstrate that people can predict their implicit attitudes (a) across a broad range of targets (many of which are plausibly difficult to infer without introspection), (b) far more accurately than third-party observers can based on demographic information, and (c) with similar accuracy across two different widely used implicit measures. On the other hand, predictive accuracy (a) varied widely across individuals and attitude targets and (b) was partially explained by inference over nonintrospective cues such as demographic variables and explicit attitudes; moreover, (c) explicit attitudes explained considerably larger portions of variance in predictions than implicit attitudes did. Taken together, these findings suggest that successful predictions of one's implicit attitudes may emerge from multiple mechanisms, including inference over nonintrospective cues and genuine introspective access.

Public Significance Statement

The terms "implicit bias" and "unconscious bias" are often used interchangeably. But are implicit attitudes truly unconscious? Past work has demonstrated that people can accurately predict the patterns of their implicit attitudes, and this finding has been interpreted as evidence that implicit attitudes are actually conscious—that is, available to introspection. However, it is also conceivable that people deduce their implicit attitudes from easily observable facts about themselves, such as their demographic categories (including age, gender, and ethnicity) and their explicit (self-reported) attitudes. Here, we find evidence for the operation of both types of processes. That is, accurate predictions of implicit attitudes seem to emerge from a mix of introspection into the contents of one's implicit attitudes and deductive reasoning relying on other, more indirect sources of evidence.

Keywords: awareness, implicit attitudes, introspection, self-perception, social cognition

Supplemental materials: https://doi.org/10.1037/xge0001464.supp

This article was published Online First September 21, 2023.
Adam Morris https://orcid.org/0000-0002-8810-8694
Benedek Kurdi https://orcid.org/0000-0001-5000-0584
Adam Morris was funded by NIH Kirschstein-NRSA postdoctoral fellowship 1F32MH131253-01A1.

Benedek Kurdi is a member of the Scientific Advisory Board of Project Implicit, a 501(c)(3) nonprofit organization and international collaborative of researchers who are interested in implicit social cognition. Preregistrations, materials, data, and analysis scripts are available via the Open Science Framework (https://osf.io/gzar7/). A previous version of this article was posted as a preprint on PsyArXiv (https://psyarxiv.com/dmjfq) and presented at the 23rd Annual Meeting of the Society for Personality and Social Psychology in February 2023 in Atlanta, Georgia.

Adam Morris and Benedek Kurdi contributed equally to conceptualization, formal analysis, funding acquisition, investigation, methodology, project administration, resources, software, supervision, validation, visualization, writing–original draft, writing–review and editing, and data curation.

Correspondence concerning this article should be addressed to Adam Morris, Department of Psychology, Princeton University, Peretsman-Scully Hall, Princeton, NJ 08540, United States, or Benedek Kurdi, Department of Psychology, University of Illinois Urbana-Champaign, 603 E Daniel Street, Champaign, IL 61820, United States. Email: adam.mtc.morris@gmail.com or kurdi@illinois.edu

A central finding in social psychology is that attitudes can be activated automatically and influence behavior in ways uniquely captured by indirect measures (such as the Implicit Association Test [IAT]; Greenwald et al., 1998). Indeed, such unintentionally revealed (or implicit) attitudes often diverge from the attitudes that people intentionally report on self-report measures (explicit attitudes; e.g., Hofmann et al., 2005; Nosek, 2005), and predict unique variance in behavior across a wide range of contexts (Greenwald et al., 2009; Kurdi et al., 2019).

A major open question about implicit attitudes is the extent to which they are conscious, or accessible to introspection (Gawronski et al., 2006; Greenwald et al., 2020). In contrast to traditional theories that defined implicit attitudes, in part, by their introspective inaccessibility (Greenwald & Banaji, 1995), a recent line of work has been taken to suggest that people may be aware of their own implicit attitudes. In this line of work, Hahn et al. (2014) asked participants to predict their scores on the IAT, a widely used measure of implicit attitudes, for five attitude targets (e.g., White people vs. Black people and children vs. adults). The authors found that people could predict their IAT scores across those targets with remarkable accuracy (see also Hahn & Gawronski, 2019; Hahn & Goedderz, 2020; Rivers & Hahn, 2019).[1] This finding has been widely interpreted as evidence that implicit attitudes are accessible to introspection, that is, that people can have direct, conscious awareness of the contents of their implicit attitudes (e.g. Cooley et al., 2015; Gawronski, 2019; Gawronski & Hahn, 2018)—with far-reaching implications for basic theoretical questions in social cognition, such as the nature of implicit attitudes (Berger, 2020) and the scope of introspection (Morris, 2021).

However, we argue here that prior work does not conclusively support this interpretation, for two reasons. First, there is a plausible alternative account of how people could be accurately predicting their implicit attitudes: Rather than becoming aware of their implicit attitudes through direct introspection, people could be inferring their IAT scores based on other information, such as knowledge of their demographic categories, knowledge of their explicit attitudes, memories of past encounters with the attitude target, or other lay theories they hold about themselves (Bem, 1972; Carruthers, 2009; Cushman, 2020; Gopnik, 1993; Nisbett & Wilson, 1977; Wilson, 2004).

To give a simple example, a 50-year-old Latino man from Miami could reason that people of his demographic are likely to have a stronger pro-Latino implicit preference than a pro-Asian implicit preference. We show below that even this simple form of demographic-based inference would be sufficient to make highly accurate predictions in Hahn et al.'s (2014) paradigm without any first-person introspective access to implicit attitudes. To the extent that predictive accuracy derives from inferential rather than introspective mechanisms, this result would undermine the evidence for introspective awareness of implicit attitudes (Bem, 1972; Nisbett & Wilson, 1977).[2]

Second, people's ability to predict their implicit attitudes has been shown only in a small subset of implicit attitudes, using only one measure of implicit attitudes, and in small college-student samples. Given these limitations, it is difficult to draw general inferences about people's awareness of their implicit attitudes (Yarkoni, 2022). This limited scope is especially troubling because the focal attitudes tested in prior work—such as racial attitudes—may be relatively easy to predict via inferential mechanisms; for instance, racial implicit attitudes are a common topic of public discourse and are known to vary across demographics (Nosek et al., 2007). Testing people's ability to predict their implicit attitudes across a wider scope of attitude objects can thus help characterize the mechanisms underlying those predictions.

Here, we address these limitations and submit people's awareness of their own implicit attitudes to more systematic and stringent tests. In large international samples, we probe whether people can predict a broad range of implicit attitudes (many of which are prima facie harder to identify through inference), assess their predictions using both an IAT and an alternative assay of implicit attitudes, and interrogate the patterns of their predictions for evidence of inferential versus introspective mechanisms. The results paint a complex picture, suggesting that people's ability to predict their implicit attitudes may stem from both genuine introspective awareness and inference based on other information. Notably, these studies are the result of an adversarial collaboration (where the two authors differed substantially in the priors that they placed on the possibility of accurate introspective access to implicit attitudes) and were thus designed to be informative to both sides of the debate.

Awareness of Implicit Attitudes

Several empirical and theoretical considerations led to the traditional idea that implicit attitudes were inaccessible to introspection (Greenwald & Banaji, 1995). For one, the attitudes revealed by indirect measures like the IAT can differ from the attitudes that people report on self-report measures (Cunningham et al., 2004; Hofmann et al., 2005; Nosek, 2005) and can predict behaviors that explicit attitudes do not (Cameron et al., 2012; Greenwald et al., 2009; Kurdi et al., 2019). Moreover, people are routinely found to express surprise (and even resistance) when receiving feedback on their IAT scores, suggesting that they were initially unaware of what their scores would be (Howell et al., 2015; Vitriol & Moskowitz, 2021). Finally, because "system-1" processes are often assumed to share a set of common properties (such as being fast, automatic, and unconscious; Evans & Stanovich, 2013), it seemed likely that the mental processes captured by an IAT—processes thought to be relatively fast and automatic (Greenwald et al., 2020)—would also operate unconsciously.

Several accounts, however, have challenged the view that implicit attitudes are unconscious. For instance, the associative–propositional evaluation (APE) model (Gawronski & Bodenhausen, 2006, 2011) posits that implicit attitudes are spontaneous affective reactions of which people could become conscious. Meanwhile, according to the APE model, explicit attitudes can incorporate additional information not captured by implicit attitudes (such as propositional reasoning). Concretely, a person might have a negative affective reaction to pictures of elderly people (which shows up as a biased

[1] Technically, Hahn et al. (2014) found that people can predict the *within-participant patterns* of their IAT scores, that is, their score for each attitude target relative to the other attitude targets. The present analyses similarly focus on people's ability to predict the within-subject patterns of their IAT scores (with some exceptions, which we note explicitly in the text). However, for brevity, we refer to this dependent variable as "predicting IAT scores"; the reader should interpret this language as predicting within-participant patterns of IAT scores (unless noted otherwise).

[2] We use the term "mechanism" here to indicate the source of information used to make predictions: Introspective mechanisms use direct, first-person conscious awareness of the contents of implicit attitudes (Dehaene, 2014; Hahn & Goedderz, 2020; Morris, 2021), whereas inferential mechanisms infer those contents from other knowledge (Nisbett & Wilson, 1977). We do not investigate the precise series of cognitive operations that transform such information into predictions.

implicit attitude on an IAT); however, when asked about their attitude explicitly, they could consider their belief in equality and their moral principle of respecting one's elders, and thus downweight their negative affective reaction and respond more positively. On this view, then, divergence between implicit and explicit attitudes can emerge not because implicit attitudes are unconscious and explicit attitudes are conscious; rather, it can emerge because explicit attitudes incorporate information or motivational elements (such as self-presentation concerns) that implicit attitudes do not (Fazio, 2007; Fazio & Olson, 2003; Gawronski & Bodenhausen, 2006). Thus, the extent to which people are aware of their implicit attitudes remains an open question.

In addition to the implications of this question for the basic nature of implicit attitudes, it also has potential moral and practical repercussions. For instance, if people are genuinely unaware of their implicit attitudes, this may reduce or alter their ethical responsibility for the harm caused by those attitudes (Banaji et al., 2015; Brownstein, 2016; Holroyd, 2012; Holroyd et al., 2017; Levy, 2014). In contrast, if people can directly introspect on their implicit attitudes, this may heighten their responsibility for controlling or changing them and may suggest new avenues for interventions to reduce unwanted implicit biases (e.g., by guiding people to introspect on their implicit attitudes directly).

Despite its importance, this question has received relatively little direct attention in empirical work. The primary evidence comes from Hahn and colleagues, who asked participants to predict their IAT scores for the following comparisons: White versus Black people, White versus Asian people, White versus Latin people, celebrities versus regular people, and children versus adults. People's predictions were remarkably accurate: When true scores were regressed on predicted scores, the standardized regression coefficient was over 0.5. The relationship remained significant when controlling for predictions about what implicit attitudes other people would show in general, suggesting that successful predictions rely on information that is specific to each participant.

Hahn et al. have replicated this finding several times, showing, for instance, that it holds even when participants are given minimal instructions (Hahn et al., 2014; Hahn & Gawronski, 2019). Moreover, using the quadruple process model (Conrey et al., 2005), Rivers and Hahn (2019) demonstrated that predictions track with the so-called "activated associations" component of IAT scores (in addition to a self-regulatory, "overcoming bias" component). This finding suggests that people are able to track the automatic evaluative aspects of their implicit attitudes, and not merely top-down control processes secondarily influencing their IAT scores. These data have been interpreted as evidence for the idea that the spontaneous affective reactions underlying IAT scores can be accessed introspectively if people are directed to attend to them (Hahn et al., 2014; Hahn & Goedderz, 2020). On this view, implicit attitudes may operate unconsciously due to inattention, but they are not permanently unconscious; their contents can enter conscious awareness via introspection.

Following Hahn et al. (2014) and Hahn and Gawronski (2019), in this article, we adopt the definition of implicit attitudes as spontaneous affective reactions, originally proposed in the APE model (Gawronski & Bodenhausen, 2006, 2011).[3] Moreover, by "awareness" or "introspective access" to implicit attitudes, we mean direct, first-order conscious experience of those spontaneous affective reactions, which renders them accessible for subsequent global cognitive processing (such as verbal report or explicit reasoning; Baars, 2005; Block, 1995; Dehaene & Naccache, 2001). In this sense, being consciously aware of (or introspectively accessing) the contents of an implicit attitude is importantly different from possessing a metarepresentation about the implicit attitude (e.g., learning from a psychology lecture that you likely have implicit racial biases; Block, 1995; Dehaene et al., 2017; Fleming et al., 2012). Intuitively, it is the difference between hearing Jingle Bells playing in your head (awareness) and thinking "Jingle Bells is playing in my head" (metarepresentation).

Although first-order awareness of an implicit attitude would likely facilitate further metacognition about the attitude, the two are importantly distinguishable and exhibit different cognitive and neural properties (Dehaene et al., 2017). Reportability of implicit attitudes is a necessary but not sufficient criterion for conscious awareness, in the sense used here; if someone can report their implicit attitudes after learning about them in a lecture, they have acquired a metarepresentation of the attitudes but have not thereby become "aware" of them in the relevant sense. Establishing genuine awareness requires understanding the mechanism underlying such reports, which is a central goal of the present project.

Open Questions About the Mechanism and Scope of Predictive Accuracy

Past work leaves open major questions about the mechanism and scope of people's ability to predict their implicit attitudes. As Hahn et al. (2014) acknowledge, the ability to predict IAT scores could be explained by an alternate, nonintrospective mechanism: People could be inferring their implicit attitudes from other knowledge about themselves, as an external observer might.

Indeed, a large body of work in social psychology suggests that people routinely make inferences about their own mind via information they observe about themselves in much the same way that a third-party observer would (Bem, 1972; Carruthers, 2009; Cushman, 2020; Gopnik, 1993; Nisbett & Wilson, 1977; Wilson, 2004). For instance, in Bem's classic work on self-perception, people routinely inferred their attitudes from observing their own choices (Bem, 1972). More recently, Bayesian models have suggested that people possess powerful cognitive mechanisms for inferring latent facts about social targets from observable information (Moutoussis et al., 2014). It is entirely plausible that people could apply these inferential tools to predict their own implicit attitudes, without any direct introspective access to those attitudes.

There is much information about the self that people could use to make these inferences, such as memories of past encounters with an attitude target. Here, we begin investigating this space by focusing on two simple types of information that could easily be used for self-inference: demographic variables and knowledge of explicit attitudes. Implicit attitudes are known to vary across demographics (Nosek et al., 2007); it is thus plausible that people's lay theories could include knowledge of this link, and that people could use

[3] Though for clarity we adopt Hahn et al.'s interpretation of implicit attitudes, in principle our studies are agnostic to how implicit attitudes are defined. Whatever constitutes the true contents of implicit attitudes, we seek to understand the mechanism and scope of people's ability to predict those contents.

information about their own demographic group memberships to predict their implicit attitudes. We demonstrate below that simple demographic variables share enough variance with IAT scores that they could be used, in principle, to predict those scores with as much accuracy as participants actually exhibit in the paradigm of Hahn et al. (2014).[4] Similarly, explicit attitudes share substantial variance with implicit attitudes, and people could be using knowledge of their own explicit attitudes to accurately predict variance in their implicit attitudes.

The possibility of inference-based predictive mechanisms has important implications. To the extent that people are predicting their IAT scores via inference rather than introspection, this would weaken the evidence for direct introspective awareness of implicit attitudes (Nisbett & Wilson, 1977) and would hark back to the traditional claim that implicit attitudes are introspectively inaccessible (Greenwald & Banaji, 1995). In contrast, to the extent that people can predict their implicit attitudes above and beyond inference over demographic information or explicit attitudes, this finding would push us one step closer to demonstrating genuine introspective awareness of implicit attitudes.

Of course, demographics and explicit attitudes are only two sources of information that people could be using to infer their own implicit attitudes and are far from the richest sources. For instance, memories of past encounters with the attitude object, or more complex lay theories about the self, could potentially provide much richer information from which to infer one's implicit attitudes. Examining the role of demographics and explicit attitudes offers a proof of concept of how inferential mechanisms could underlie implicit attitude predictions, and a tractable first step in trying to adjudicate between inferential and introspective mechanisms.

A second major open question concerns the scope of people's predictive accuracy. Hahn et al. only tested five, relatively homogenous attitude comparisons (White vs. Black people, White vs. Asian people, White vs. Latin people, celebrities vs. regular people, and children vs. adults). However, the range of implicit attitude targets is, of course, much broader, spanning in principle the entire range of concrete stimuli (or even abstract concepts) that humans encounter in their daily lives. Moreover, these implicit attitudes vary widely in their properties, such as the extent to which they diverge from explicit attitudes or elicit self-presentation concerns (Hofmann et al., 2005; Nosek, 2005). A key open question is the extent to which predictive accuracy generalizes beyond the small set of targets used in prior work.

Critically, the limited scope of attitude targets used in existing studies exacerbates the mechanistic uncertainty described above. The five targets tested in Hahn et al. (2014) are widely present in public discourse and, plausibly, in people's lay theories about themselves. Attitudes about these targets may therefore be especially easy to infer from nonintrospective mechanisms. In contrast, attitudes toward other targets (such as some of the abstract ones tested in the present work, including emotions vs. reason or approach vs. avoid) may be prima facie more difficult to infer. Testing predictive accuracy across a broad range of attitudes, then, also helps adjudicate between introspection-based and inference-based mechanisms underlying predictions.

A related open question is whether predictive accuracy generalizes across measures of implicit attitudes. Although the IAT has been the most widely used individual-difference measure of implicit attitudes, it has some unique properties that represent only one of many possible instantiations of the latent construct it is intended to measure. For instance, the proposed mechanism of the IAT is response competition (De Houwer et al., 2009; De Houwer & Moors, 2010), which is only one of the several ways in which the suboptimal conditions thought to be crucial to implicit attitude measurement can be instantiated. In addition, unlike most other implicit measures, the IAT requires categorization of the attitude objects, thus making an IAT a so-called "relevant-feature paradigm" (De Houwer, 2003; Field et al., 2011)

A popular alternative measure—the affect misattribution procedure (AMP; Payne et al., 2005)—measures implicit attitudes via the misattribution of affect, does not require categorization of the attitude objects (thus making it an "irrelevant-feature paradigm," De Houwer, 2003; Field et al., 2011), and has been shown to capture variation in implicit attitudes that is relatively independent of what is measured by the IAT (Bar-Anan & Vianello, 2018; Van Dessel et al., 2019). If people can also predict their attitudes as measured by the AMP, this would suggest that people are aware of broader, underlying implicit constructs that span the two measures. In contrast, if predictions track only the IAT, this would suggest that predictive accuracy is narrower and may be linked to some aspect of IAT performance that does not generalize beyond this particular measure.[5]

Finally, people's ability to predict their implicit attitudes has primarily been tested in samples of college students (Hahn et al., 2014; Hahn & Gawronski, 2019). It is unknown how well predictive accuracy observed in this participant group would generalize to more diverse and more representative samples. For instance, prior work may overestimate the extent to which people can generally predict their implicit attitudes because college students are specifically trained to notice their biases or introspect on their own attitudes (e.g., in the introductory psychology classes that they take). In the present studies, the use of large online samples allows us to investigate this possibility.

[4] To exhibit accuracy in this paradigm, people would only need to know their own demographic categories and the patterns of spontaneous affective reactions likely associated with those demographics. For example, a Latino man from Miami would merely have to believe that Latino Miamians are likely to have a stronger pro-Latino implicit preference than a pro-Asian implicit preference. Given the complexity and power of people's lay theories about themselves and their demographic groups (Argyle, 2013; Tajfel, 2010), it is highly plausible that people possess this knowledge. A potential concern with the demographic inference account is that people often overemphasize the degree to which they are individually unique (Alicke & Govorun, 2005; Pronin & Kugler, 2010), and hence may be unlikely to use group based information (such as demographic categories) to predict their implicit attitudes. However, implicit attitudes are a relatively novel construct for participants, potentially making them more willing to believe that they are predictable via demographic categories. Moreover, people often identify strongly with their demographic identities and believe that they are similar to other in-group members (Hogg, 2016; Tajfel, 2010). Thus, we believe it is plausible that people use demographic information to infer their implicit attitudes. (We thank Jan De Houwer for raising this concern.).

[5] Hahn et al. (2014) showed that people can predict their IAT scores even when not given a detailed description of the IAT (and instead just asked to predict their implicit attitudes in more general, conceptual terms). This finding helps rule out that, mechanistically, people are predicting their implicit attitudes by simulating themselves performing an IAT-like task. Nonetheless, it is still an open question whether the variance they predict is uniquely related to the variance measured by the IAT, or whether people are tapping into awareness of a construct that spans implicit measures.

The Present Studies

To address these issues and advance our understanding of the mechanism and scope of implicit attitude prediction, we start by replicating Hahn et al.'s result with five attitude targets in two large, online samples (Studies 1A and 1B). Then, we test whether people can predict a considerably broader range of 57 implicit attitudes, measuring predictive accuracy with both the IAT (Studies 2A–2C) and the AMP (Studies 3A–3B). Critically, this large set of attitude targets allows us to begin to interrogate the mechanisms underlying predictions—by testing, for instance, whether targets attitudes toward which can be easily inferred elicit more accurate predictions, and whether people can still predict implicit attitudes for targets attitudes toward which are more difficult to infer. Finally, following the methodology of Nisbett and Wilson (1977), we recruit a sample of observers to predict the implicit attitudes of the initial participants from Study 2A based solely on information about their demographic group membership (Study 4), and compare people's predictions for themselves with those of third-party observers.

Studies 1A and 1B

Studies 1A and 1B are close replications of the paradigm from Hahn et al. (2014) and Hahn and Gawronski (2019). In this paradigm, participants are introduced to the idea of implicit attitudes (which are described as spontaneous affective reactions to target stimuli) and asked to predict what their implicit attitudes will show for five target comparisons: White people versus Black people, White people versus Asian people, White people versus Latin people, celebrities versus regular people, and children versus adults. Then, participants take all five corresponding IATs, thus allowing for a comparison of their predictions to their IAT scores.

Despite the far-reaching influence of Hahn et al.'s results, to our knowledge, they have never been replicated by an independent laboratory, and have been obtained exclusively from in-person college-student samples. In these replications, we test whether the findings of predictive accuracy generalize to a more diverse sample recruited online. All instructions, stimuli, and statistical analyses closely mirrored those from Hahn et al. As Studies 1A and 1B are nearly identical, we report them together (noting the few aspects in which they differ from each other).

Method

Transparency and Openness

For all studies, we report how we determined the sample size as well as all data exclusions, manipulations, and measures. Across all studies, we recruited the largest possible sample without overexerting the pool of volunteer participants available via the Project Implicit educational website. No formal power analyses were conducted. In all studies, sample sizes were determined a priori, without intermittent data analyses. Sample sizes, exclusion criteria, designs, and analysis plans were formally preregistered for all studies. All preregistrations, materials, stimuli, data, and analysis scripts are available from the Open Science Framework (OSF; https://osf.io/gzar7/).

Participants and Design

All studies were approved by the Yale University Institutional Review Board and conform to American Psychological Association ethical standards. A total sample of 632 volunteer participants were recruited in Study 1A and a total sample of 549 volunteer participants were recruited in Study 1B, both from the Project Implicit educational website (https://implicit.harvard.edu/). As preregistered, participants were excluded from analyses if they (a) did not complete all five prediction items ($n = 10$ in Study 1A and $n = 3$ in Study 1B), (b) did not complete all five IATs ($n = 132$ in Study 1A and $n = 53$ in Study 1B), or (c) produced response latencies of 300 ms or less on at least 10% of trials on any of the IATs, indicating careless responding ($n = 6$ in Study 1A and $n = 13$ in Study 1B).

Following these exclusions, the final sample size was 484 in Study 1A (294 women, 181 men, six participants of other genders, and three participants with missing data on the gender item; $M_{age} = 37$ years, $SD = 15$) and 480 in Study 1B (292 women, 168 men, 17 participants of other genders, and three participants with missing data on the gender item; $M_{age} = 35$ years, $SD = 16$). In both studies, participants were recruited without regard to country of origin, but the majority of participants ($n = 347$ or 72% of the final sample in Study 1A and $n = 365$ or 76% of the final sample in Study 1B) were from the United States. Additional demographic variables, including state of residence (U.S. participants only), highest educational attainment, occupation, political identification, race/ethnicity, religion, and religious identification, are available in the open data. For details of how all demographic variables were collected, see the OSF repository.

Participants in both studies read initial instructions, responded to five prediction items, completed the five corresponding abbreviated IATs (Greenwald et al., 1998), and finally answered self-report items measuring explicit attitudes and prior experience completing IATs.

Materials

IAT Category Stimuli. The IAT category stimuli were 10 color images for each category, borrowed from Hahn et al. (2014) and Hahn and Gawronski (2019). These images were originally obtained from a normed stimulus set by Minear and Park (2004) and from online searches. Specifically, for each IAT, the category stimuli included five unique facial images of men and five unique facial images of women for each category. The category labels changed from IAT to IAT depending on the social group comparison used.

IAT Attribute Stimuli. The IAT attribute stimuli were 10 words each, borrowed from Hahn and Gawronski (2019). The positive attribute stimuli included BEAUTIFUL, CHEERFUL, DELIGHT, ENJOY, HELPFUL, JOY, POSITIVE, SMILE, SUCCESS, and WONDERFUL (presented in all caps), and the negative attribute stimuli included ANGRY, BRUTAL, DESTROY, DISASTER, EVIL, HATE, HORRIBLE, TERRIBLE, TRAGIC, and UGLY (also presented in all caps). The attribute labels were "good" and "bad."

Procedure and Measures

Initial Instructions. The initial instructions in Study 1A informed participants about the existence of explicit attitudes (i.e., attitudes that one consciously endorses and is willing and able to report) and implicit attitudes (i.e., gut reactions that one feels when encountering a stimulus). Furthermore, participants were told that these two types of attitudes can be in line with each other but that they can also diverge, and that there was a test called the IAT that measured implicit attitudes. Finally, they were told that they would be asked to predict their IAT scores for five tests.

In Study 1B, we used a different set of initial instructions that followed more closely the procedure of Hahn and Gawronski (2019). We used these instructions because it was expected that they would maximize the strength of relationship between the prediction items and implicit attitudes (A. Hahn, personal communication, March 29, 2021). However, although these instructions were more detailed than the instructions used in Study 1A, their content was highly similar. In addition, in Study 1B, following the initial instructions, participants newly completed a practice prediction item, which had the same format as the actual prediction items completed subsequently but used cats and dogs (two attitude objects not included in the IATs) as examples. Although Study 1B closely paralleled the instructions and procedure of Hahn and Gawronski (2019), there remained a number of minor differences, such as font sizes, the location of pictures on the prediction item screen, etc. For completeness, we enumerate these differences in the online supplemental materials.

Prediction Items. Participants completed five prediction items (Hahn et al., 2014; Hahn & Gawronski, 2019), each corresponding to one of the subsequent IATs, in individually randomized order.

In Study 1A, the prediction item read "I predict that the IAT comparing my gut reactions to (attitude object 1 [e.g., Asian people]) and (Attitude Object 2 [e.g., White people]) will show that…." Participants responded to this item using a 100-point sliding scale. The scale was labeled, in equidistant units starting from the left endpoint of the scale and ending at the right endpoint of the scale, "a lot more positive toward (Attitude Object 1 [e.g., Asian people])," "moderately more positive toward (Attitude Object 1 [e.g., Asian people])," "slightly more positive toward (Attitude Object 1 [e.g., Asian people])," "same" (at the midpoint of the scale), "slightly more positive toward (Attitude Object 2 [e.g., White people])," "moderately more positive toward (Attitude Object 2 [e.g., White people])," and "a lot more positive toward (Attitude Object 2 [e.g., White people])." This item was scored from −50 (corresponding to the left endpoint of the scale) to +50 (corresponding to the right endpoint of the scale).

In Study 1B, the text and scoring of the prediction item remained the same. However, again in an attempt to maximize the relationship between the prediction items and implicit attitudes, the images representing each category and used on the subsequent IAT were presented in two rows, below one another (A. Hahn, personal communication, March 29, 2021). In addition, the following sentences were presented above each prediction item: "Please look at these pictures. All pictures in the top row belong to the category (Attitude Object 1 [e.g., Asian people]), and all pictures in the bottom row belong to the category (Attitude Object 2 [e.g., White people]). Listen to your gut reactions while you look at the pictures. What is your implicit attitude toward these categories? What will the IAT show?"

Implicit Attitudes. Implicit attitudes were measured using an abbreviated version of the IAT (Greenwald et al., 1998) that did not include any practice blocks (see also Cunningham et al., 2001, 2004; Phelps et al., 2000, 2003). We provide additional evidence for the validity of this approach in Studies 2A–2C below. This version of the IAT was implemented to prevent participant fatigue and to ensure that each participant had sufficient time to complete all five tests. The five IATs were completed in the same randomized order as the preceding prediction items.

Each IAT consisted of two blocks, with 40 trials each. In the first block, participants used the E key to sort stimuli belonging to one category (adults on the adult/child IAT, regular people on the regular people/celebrities IAT, and White people on the White/Asian, White/Black and White/Latino IATs) and positive attribute words to the left and the I key to sort stimuli belonging to the other category (children on the adult/child IAT, celebrities on the regular people/celebrities IAT, Asian people on the White/Asian IAT, Black people on the White/Black IAT, and Latino people on the White/Latino IAT) and negative attribute words to the right. In the second block, the assignment of categories to sides was reversed such that children, celebrities, Asian people, Black people, and Latino people were sorted together with positive attribute words and adults, regular people, and White people were sorted together with negative attribute words. Similar to Hahn et al. (2014) and Hahn and Gawronski (2019), the order of the two blocks was not counterbalanced to eliminate this irrelevant source of variance in IAT scores.

Performance on each IAT was scored using the improved scoring algorithm (Greenwald et al., 2003) such that higher IAT scores correspond to a preference for the first over the second attitude object.

Explicit Attitudes. Explicit attitudes toward adults, Asian people, Black people, celebrities, children, Latino people, regular people, and White people were measured using a feeling thermometer item each. These items were presented in individually randomized order. The text of the item read, "How coldly or warmly do you feel toward (attitude object [e.g., Asian people])?" and participants entered their response using 100-point sliding scales whose left endpoint, midpoint, and right endpoint were labeled "extremely coldly," "neither coldly nor warmly," and "extremely warmly," respectively. These explicit attitude items were then used to calculate five explicit attitude difference scores to parallel the prediction scores and IAT D scores.

Prior Experience. For each IAT, participants were asked to report whether they had completed that specific IAT prior to the study. The response options for this item included yes, no, and unsure. In addition, participants were also asked to report the total number of IATs that they had completed before the study. The response options for this item included 0, 1–5, 6–10, more than 10, and unsure.

Analytic Strategy

Standardizing Variables. When testing implicit attitude predictions, a fundamental analytical question is whether to standardize the variables at the participant level. Hahn et al. (2014) argued that all variables—that is, prediction scores and IAT D scores—should be standardized at the participant level. This approach tests participants' ability to predict where their implicit attitude scores fall relative to their other implicit attitudes (e.g., whether their anti-Black bias is bigger or smaller than their anti-Asian bias), rather than their ability to predict where their scores fall in an absolute sense within the population. Hahn et al. argue that the former is the appropriate test of introspective awareness (Hahn et al., 2014; Hahn & Goedderz, 2020); the latter, in addition to requiring knowledge of one's own implicit attitudes, also requires social knowledge not attainable through introspection (i.e., the distribution of other people's implicit attitudes), and thus could artificially deflate participants' apparent predictive accuracy. Hahn et al. refer to this additional social knowledge required to predict unstandardized D scores as "social calibration" (Hahn & Goedderz, 2020).

Following this logic, for our primary analyses we standardize all variables at the participant level. The only exception is item-level

analyses, which require unstandardized variables; we note and justify these deviations below. Nonetheless, in the present data we find no difference between standardized and unstandardized analyses (see the online supplemental materials for details and discussion). Hence, the present data do not provide evidence for the distinction between awareness and social calibration.

Testing Predictive Accuracy. In Model 1, IAT D scores were predicted in a mixed-effects model with the prediction score as the sole fixed effect and a prediction score random slope for each participant and each IAT comparison (adults/children, regular people/celebrities, White people/Asian people, White people/Black people, and White people/Latino people). Participant-level random intercepts were omitted because the data were standardized to each participant, but item-level random intercepts were included. This model provides an initial test of whether participants can accurately predict their implicit attitude scores.

It should be noted that Hahn et al. (2014) did not include item-level random effects in their models; in the online supplemental materials, we reestimate all the models without item-level random effects and show that none of the analyses reported here depend on this analytic choice, with one minor exception noted in the text below. However, by including item-level random effects, we may slightly underestimate participants' zero-order predictive accuracy in Studies 1A–1B (see the online supplemental materials for discussion.)

In Model 2, following Hahn et al. (2014), we additionally included a main effect for explicit attitudes. One deflationary account of people's predictive accuracy is that they are simply predicting the portion of their implicit attitudes that overlaps with their explicit attitudes. Concretely, since IAT scores tend to correlate with explicit attitudes, participants could make above-chance predictions of their IAT scores by simply recapitulating their explicit attitudes. By controlling for explicit attitudes, Model 2 helps rule out this deflationary account and tests whether people can predict variance in their implicit attitudes that is not captured by explicit attitudes. We refer to this variance as the "uniquely implicit" component of predictive accuracy.

We additionally test whether adding explicit attitudes as a regressor in Model 2 significantly reduces the coefficient associated with the prediction item. To accomplish this, we bootstrap 95% confidence intervals [CIs] around the difference in the target coefficient between Model 1 and Model 2 (Mooney et al., 1993). Specifically, we (a) resample participants with replacement until we have created a new sample of the same size as the original; (b) compute the difference between coefficients in that bootstrapped sample; (c) repeat this procedure 1,000 times; and then, (d) using the "bca" function from the R package *coxed* (Kropko & Jeffrey, 2019), use those vector of differences to compute a 95% bias-corrected, accelerated CI (DiCiccio & Efron, 1996). If this CI does not contain zero, we consider the reduction in the size of the coefficient from Model 1 to Model 2 to be significant.

Another deflationary account is that people could obtain above-chance predictions simply by predicting what the general population would, on average, show on each IAT comparison. For instance, if someone knows (or guesses correctly) that the general population is, on average, more biased against Black people than against Asian people, she could predict a stronger anti-Black than anti-Asian implicit bias and achieve above-chance accuracy without any individual-specific information about herself. To rule out this alternative, following Hahn et al. (2014), we conducted a permutation test. In Hahn et al. (2014), the authors randomly paired each participant's predictions with a different participant's IAT scores and found that participants' predictions corresponded to their own IAT scores more than a random other person's scores.

Here, we adopted a similar procedure, except instead of one random pairing, we created 1,000 random pairings. For each random pairing of predictions and actual IAT scores, we calculated the t statistic obtained from regressing scores on predictions (using the mixed-effects model described above), resulting in a distribution of t statistics under the null hypothesis that participants are simply predicting the general population's average IAT scores (with no individual-specific information about themselves). We then derived a p value by calculating the proportion of the null distribution greater than the empirical t statistic obtained using the actual pairings between participants' predictions and scores. This p value expresses the probability that the predictive accuracy observed in the empirical sample was due to some aspect of responding shared across (rather than specific to) participants.

In all mixed-effects models, to test for overparameterization, we followed the procedure recommended by Bates et al. (2018) and Matuschek et al. (2017) and conducted a principal component analysis on the random effects covariance matrix of the fitted model. If we found that there were principal components accounting for 0% of the variance (after rounding to three decimal places), we simplified the model in two steps. In the first step, we fixed the correlation between the item-level random intercepts and slopes to zero.[6] If this step did not resolve the issue, in a second step we removed the random slope for whichever level (participants and/or items) had a degenerate random effects covariance matrix. Note that this procedure may result in different final random effects structure across models. For succinctness, we do not report the final structure of each model in the main text; however, all model structures are reported in the online supplemental materials.

Demographic Prediction Analysis. Finally, to probe whether predictive accuracy could in principle be the result of a simple inferential mechanism—participants inferring their IAT scores from their demographics—we investigated how accurately participants' IAT D scores could be predicted based on their demographic information alone. For each of the five IAT target comparisons (e.g., Black vs. White, adult vs. child), we regressed participants' unstandardized IAT scores on demographic variables, including age, gender, race/ethnicity, education level, political orientation, and religiosity.[7] In each regression, to prevent overfitting, we left out one participant, and then used the estimated regression model to predict the left-out participant's IAT score from that participant's demographics.

We repeated this procedure for each participant and each IAT comparison until we had obtained demographically predicted IAT scores for all five comparisons for each participant. Finally, we regressed participants' actual IAT scores on these demographically predicted IAT scores using the same mixed-effects model as

[6] In the preregistrations for Studies 1A–B, we accidentally omitted this step, and instead said that we would simplify the models by immediately removing the random slopes. The step of fixing the correlation between random intercepts and slopes to zero is included in the preregistrations of all the remaining studies. None of the results in Studies 1A–B change if we omit this step in the analysis.

[7] We omitted additional demographic variables such as nationality because they were too high-dimensional to include in the regression.

described above. This analysis—which we refer to as the "demographic prediction analysis"—reveals the extent to which IAT scores can in principle be predicted from demographic information. (This analysis was not preregistered; we conceived of it after conducting the studies. It should therefore be viewed as exploratory.)

Note that the demographic prediction analysis requires us to use participants' unstandardized IAT scores, rather than standardizing the scores within each participant as we do for other analyses. The demographic prediction analysis is an item-level (attitude target-level) analysis: It uses the relationship between other participants' demographics and implicit attitudes for a particular attitude target to predict the D score of a left-out participant for that attitude target. Standardizing the variables would lead item-level analyses like this one awry by changing each participant's D scores for a particular attitude target (e.g., Black vs. White) relative to their scores for the other attitude targets not currently being examined. Leaving the variables unstandardized is thus required for item-level analyses. (For these analyses, we still report standardized regression coefficients, but these are standardized across the whole sample, not at the participant level.)

A caveat of using unstandardized variables is that, per Hahn et al.'s (2014) logic above, the resulting findings may relate to social calibration in addition to introspective awareness. Concretely, it may be that part of what demographic information can predict is the absolute position of people's D scores within the population, rather than just the within-participant pattern of D scores. However, note that predicting unstandardized D scores would theoretically require both awareness (i.e., knowledge of one's own implicit attitudes) and social calibration (knowing where those scores fall in the population; Hahn & Goedderz, 2020). Hence, if demographic information can predict unstandardized D scores, this finding would still indicate that implicit attitudes can be inferred from demographics.[8] Moreover, as reported in the online supplemental materials, we find little difference in the present studies between standardized and unstandardized analyses, suggesting that awareness and social calibration are not distinct in these data.

Computational Tools. Statistical analyses for all studies were performed in the R statistical computing environment (Version 4.0.3); mixed-effects models were fit using the *lme4* (Bates et al., 2015) and *lmerTest* (Kuznetsova et al., 2017) packages.

Results

Descriptive Statistics

The relationship between the prediction item and implicit attitudes in the overall sample and at the participant level is shown in Figures 1 (Study 1A) and 2 (Study 1B).

In Study 1A, participants' predictions were positively correlated with their IAT D scores across the whole sample ($r = .096$). The median participant-level correlation was $r = .110$ ($SD = 0.531$). In Study 1B, the prediction item was also positively correlated with IAT D scores in the whole sample ($r = .249$). The median participant-level correlation was $r = .325$ ($SD = 0.525$).

Interestingly, in both studies, there was considerable participant-level variability in predictive accuracy, with many participants (45% in Study 1A and 30% in Study 1B) showing a negative correlation between their predicted and actual scores. We do not examine this participant-level variability in detail in these studies but return to this issue in the General Discussion section.

Mixed-Effects Models

In Model 1, which used the prediction items as the only independent variable, we obtained a small but statistically significant relationship between the prediction items and IAT D scores in both Study 1A, $\beta = .160$, $t(3.01) = 5.92$, $p = .010$, and in Study 1B, $\beta = .157$, $t(569.47) = 6.54$, $p < .001$. Based on the results of the permutation test, the significant relationship was specific to individual participants and could not be accounted for by population-level associations in either study ($ps < .001$). (Note that, in Study 1B, the prediction coefficient increased to 0.25 when excluding item-level random effects; see the online supplemental materials for details and discussion.)

In Model 2, we found that the relationship between the prediction item and implicit attitudes persisted following the addition of the explicit attitude difference scores to the model. In both studies, the effect size associated with the prediction item was reduced but remained significant—Study 1A, $\beta = .130$, $t(4.42) = 4.35$, $p = .010$; Study 1B, $\beta = .105$, $t(705.71) = 4.04$, $p < .001$—and explicit attitudes also produced a significant effect—Study 1A, $\beta = .084$, $t(2{,}273.05) = 3.51$, $p < .001$; Study 1B, $\beta = .122$, $t(2{,}259.61) = 5.22$, $p < .001$. The reduction in the coefficient for the prediction item was significant; the bootstrapped 95% CIs around the difference in coefficients between Model 1 and Model 2 did not contain zero (Study 1A, [0.009, 0.054], Study 1B: [0.028, 0.078]).

Note that Model 2 in Studies 1A–1B was the only model in the article that produced qualitatively different results when excluding item-level random effects (the analytic approach adopted by Hahn et al., 2014, in their original article). When excluding item-level random effects, the coefficient for explicit attitudes was not significant—Study 1A, $\beta = .021$, $t(2{,}311.42) = 0.87$, $p = .386$; Study 1B, $\beta = .001$, $t(374.31) = 0.04$, $p = .965$—and adding explicit attitudes to the model did not significantly reduce the prediction item coefficient (95% CIs for Study 1A, [−0.021, 0.035], for Study 1B: [−0.027, 0.018]). Thus, we do not interpret these findings strongly for Studies 1A–1B.

Subgroup Analyses

We obtained the same pattern of results when analyses were restricted to participants who were (a) naïve with respect to each specific IAT or (b) fully naïve (i.e., had never completed an IAT prior to the study). These subgroup analyses did not differ from the full-group analyses in any of our subsequent studies; hence, they are not reported in the article for this or any of the remaining studies but are available in the open code.

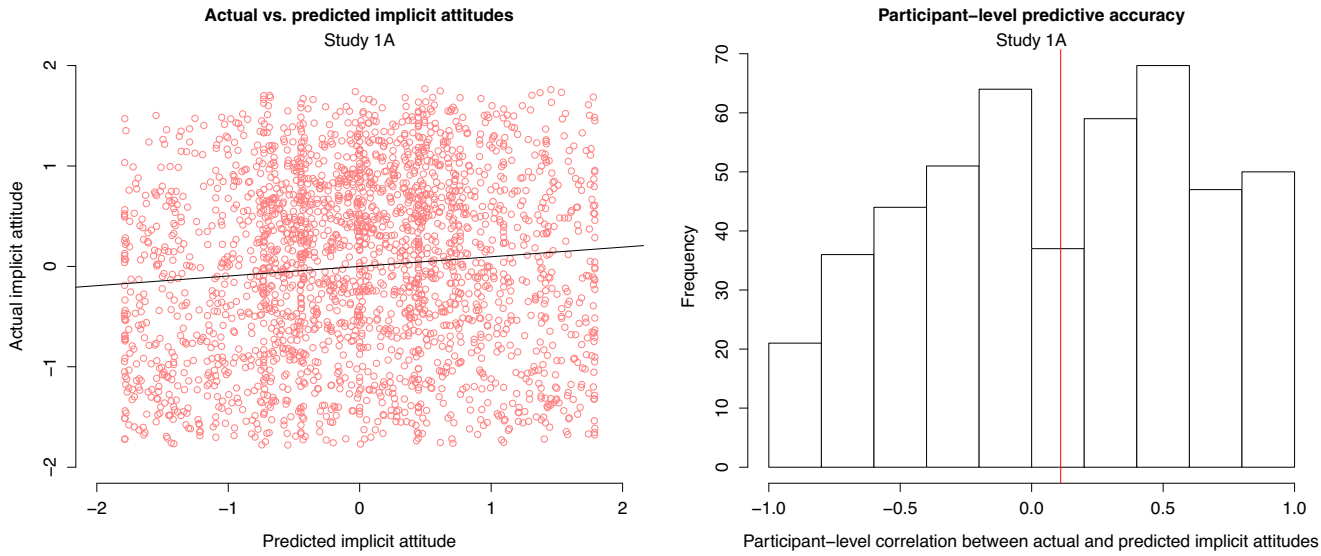
Predicting IAT Scores From Demographic Information

The scores obtained via the leave-one-out demographic prediction analysis significantly predicted IAT scores, $\beta = .215$, $t(5.56) = 6.83$, $p < .001$, in Study 1A, and $\beta = .189$, $t(3.79) = 3.40$, $p = .030$, in Study 1B, with similar levels of accuracy to participants' predictions.

[8] To put it another way: If someone could infer their unstandardized D scores from knowledge of their demographics, all they would have to do to obtain their standardized D scores is standardize them. Predicting unstandardized variables in this context is a harder feat than predicting standardized ones (Hahn et al., 2014; Hahn & Goedderz, 2020).

Figure 1

Overall Correlation Between the Standardized Prediction Item and Standardized Implicit Attitudes (Left Panel) and Distribution of Participant-Level Correlations Between the Standardized Prediction Item and Standardized Implicit Attitudes (Right Panel) in Study 1A



Note. See the online article for the color version of this figure.

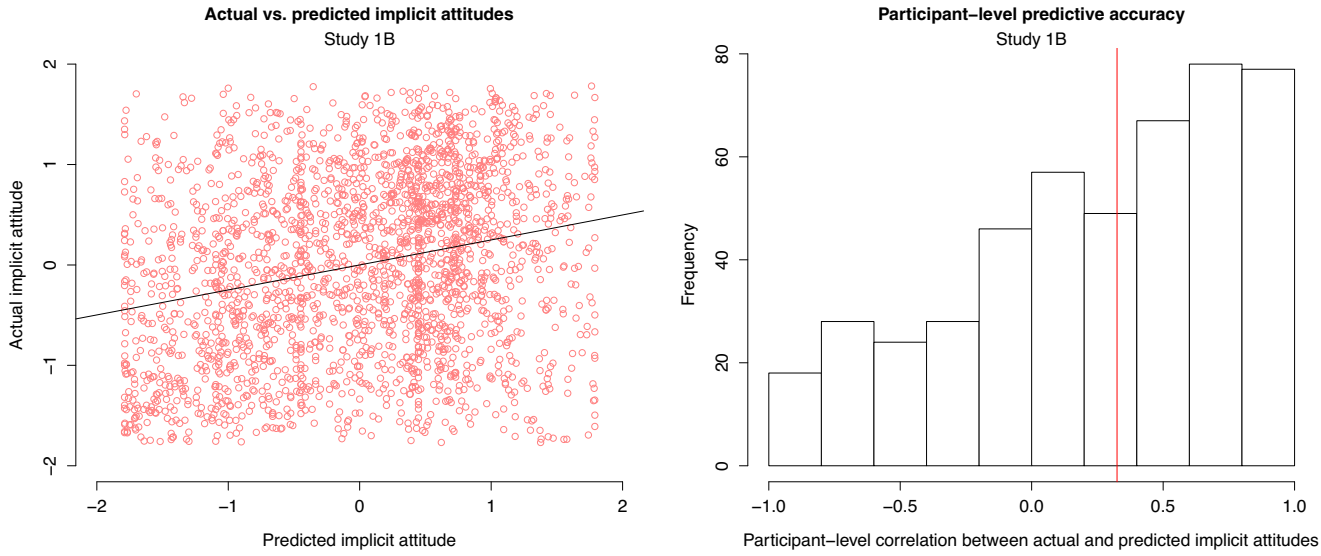
Discussion

Using a large online sample, Studies 1A–B replicated the basic finding from Hahn et al. (2014): Participants could predict their own IAT scores above chance for the five comparisons used (White vs. Black people, White vs. Asian people, White vs. Latin people, celebrities vs. regular people, and children vs. adults). Moreover, predictions were still above chance when controlling for explicit attitudes, suggesting that predictive accuracy could not be explained by participants merely recapitulating their explicit attitudes. These findings underscore the robustness and generalizability of the results obtained by Hahn and colleagues across samples (relatively homogeneous student sample vs. relatively diverse nonstudent sample) and settings (in lab vs. online).

However, the present results diverge from prior findings in two ways. First, participants exhibited considerably lower levels of

Figure 2

Overall Correlation Between the Standardized Prediction Item and Standardized Implicit Attitudes (Left Panel) and Distribution of Participant-Level Correlations Between the Standardized Prediction Item and Standardized Implicit Attitudes (Right Panel) in Study 1B



Note. See the online article for the color version of this figure.

accuracy than found in Hahn et al. (2014) and Hahn and Gawronski (2019). Hahn et al. reported standardized regression coefficients of 0.4–0.5; here, we find coefficients of 0.16 (or 0.25, if excluding item-level random effects)—a twofold or threefold decrease. Moreover, we observed considerable variation in predictive accuracy across participants, with over a third of participants exhibiting negative correlations between their predicted and actual IAT scores. These findings suggest that participants' predictive accuracy may be more variable than previously assumed on the basis of student samples alone. There are several potential explanations for these differences—most notably, the fact that the present studies were run online, whereas the studies by Hahn and colleagues had been conducted in person. We explore potential causes of the discrepancy in effect sizes, and potential sources of between-participant variability, in the General Discussion section.

Second, in the demographic prediction analysis, we found that participants' (unstandardized) IAT scores could in principle be predicted by their demographic information alone, with similar levels of accuracy as participants exhibit when making their predictions. Of course, the fact that IAT scores could be predicted from demographic information does not mean they actually are. However, this analysis highlights the need to probe whether inferential mechanisms (such as inferring one's likely implicit attitudes from one's demographics) may, at least in part, underlie people's ability to predict their IAT scores.

Study 2A

In Study 2A, we test whether participants can predict their IAT scores across a broad and diverse set of 57 implicit attitude targets, drawn from Nosek (2005). This set includes many common attitude targets such as racial or political groups but also features less usual comparisons such as emotions versus reason and approach versus avoid. The large set of attitude targets used in Study 2A allows us to shift focus from *whether* people can predict their implicit attitudes (the way the question has typically been posed in past work) to *when* they are able to do so: Which implicit attitudes can people predict well, and which are more inscrutable? By probing the sources of between-target variability in predictive accuracy, Study 2A constitutes a first step toward gaining a more mechanistic understanding of how successful predictions of implicit attitudes come about.

More specifically, testing predictive accuracy across this broader range of attitudes can help accomplish several goals. Since past work relied on a small set of relatively homogeneous attitude targets, the present study can provide information on whether predictive accuracy generalizes across a broader space of attitude objects. This test is intrinsically important for understanding the scope and generalizability of successful implicit attitude prediction. Indeed, as explained in the introduction, people might plausibly be able to predict the small set of attitudes investigated in past work without being aware of their implicit attitudes more broadly.

Moreover, by including a relatively large set of attitude objects, Study 2A allows for a quantitative investigation of target-level predictors of predictive accuracy. Indeed, we probe people's ability to predict implicit attitudes that are plausibly harder to identify through nonintrospective, inferential mechanisms, including attitudes that cannot be predicted from demographic information alone, and which exhibit low correlation between implicit and explicit measures. If predictive accuracy is undiminished for such attitudes, this finding would be consistent with an introspective mechanism. In contrast, if people can better predict the attitudes that are more easily predicted via demographics or explicit attitudes, this finding would provide evidence that inference over these information sources plays a role in successful predictions.

Method

Participants and Design

A total sample of 2,241 volunteer participants were recruited from Project Implicit (https://implicit.harvard.edu/). As preregistered, participants were excluded from analyses if they (a) did not complete all five prediction items ($n = 19$), (b) did not complete all five IATs ($n = 275$), or (c) produced response latencies of 300 ms or less on at least 10% of trials on any of the IATs, indicating careless responding ($n = 104$).

Following these exclusions, the final sample size was 1,843 (1,191 women, 608 men, 27 participants of other genders, and 17 participants with missing data on the gender item; $M_{age} = 38$ years, $SD = 16$). Participants were recruited without regard to country of origin, but the majority of participants ($n = 1{,}350$ or 73% of the final sample) were from the United States.

The design was the same as in Studies 1A and 1B. That is, participants read initial instructions, responded to five prediction items, completed the five corresponding abbreviated IATs (Greenwald et al., 1998), and finally answered self-report items measuring explicit attitudes and prior experience completing IATs. However, unlike in Studies 1A and 1B, where all participants completed the same five IATs, in this study participants completed a subset of five IATs randomly selected from a larger set of 57 (Nosek, 2005).

Materials

IAT Category Stimuli. For 16 of the 57 IATs, the category stimuli were six color images representing each category. For the remaining IATs, the category stimuli were six words representing each category. For example, on the image-based Coke/Pepsi IAT, the category stimuli were six images related to each brand. On the word-based U.S./Japan IAT, the category stimuli were six words related to each country ("Washington," "New York," "Mt. Rushmore," "Chrysler," "Ford," and "America" vs. "Tokyo," "Kyoto," "Mt. Fuji," "Mitsubishi," "Honda," and "Asia"). The images were obtained using online searches, whereas the words were adapted from Nosek (2005). The category labels changed from test to test, depending on the specific comparison used.

Given that the number of category stimuli differed across IATs in Nosek (2005), for some IATs, new category stimuli were added and for other IATs, some category stimuli were omitted. In addition, three IATs from Nosek (2005) were deemed to be dated and were therefore fully replaced: the Al Gore/George W. Bush IAT with a Joe Biden/Donald Trump IAT, the Jay Leno/David Letterman IAT with a Jimmy Fallon/Stephen Colbert IAT, and the Meg Ryan/Julia Roberts IAT with a Nicole Kidman/Reese Witherspoon IAT.

IAT Attribute Stimuli. The IAT attribute stimuli were six words each. The positive attribute stimuli included "good," "great," "fantastic," "pleasant," "positive," and "wonderful," and the negative attribute stimuli included "awful," "bad," "horrible," "negative," "terrible," and "unpleasant." The attribute labels were "good" and "bad."

Procedure and Measures

The initial instructions, prediction items, IATs, and prior experience items were the same as in Study 1B. However, unlike in Study 1B, participants completed two practice prediction items, one for an image-based IAT (insects vs. flowers) and one for a word-based IAT (nature vs. civilization). In addition, as mentioned above, participants completed a random subset of five IATs, drawn from a larger set of 57, rather than the same five IATs.

Finally, given that feeling thermometer measures of explicit attitudes tend to be particularly highly correlated with IATs (Hofmann et al., 2005), the unique contribution of prediction items may have been underestimated in Studies 1A–1B. As such, to obtain a more conservative estimate, in this study and all remaining studies we used a liking measure of explicit attitudes rather than a feeling thermometer. Specifically, participants were asked to indicate how much they disliked or liked each attitude object using a 100-point sliding scale. The endpoints of the scale were marked *dislike a lot* and *like a lot*, respectively, with the midpoint labeled *Neither dislike nor like*. We revisit the issue of differences and similarities across explicit attitude measures more directly in Study 2C.

Results

Descriptive Statistics

The relationship between the prediction item and implicit attitudes in the overall sample and at the participant level is shown in Figure 3. Similar to Studies 1A and 1B, the prediction item was positively correlated with IAT D scores in the whole sample ($r = .369$). The median participant-level correlation was $r = .480$ ($SD = 0.467$). Descriptively, participants' predictions were substantially more accurate in Study 2A than in the two previous studies.

Validating the Abbreviated IATs

By using Nosek's (2005) set of attitude targets, Study 2A gave us the opportunity to examine whether the abbreviated IATs produced similar results as the corresponding full-length IATs. Specifically, we tested whether participants showed similar patterns of implicit attitudes across the attitude targets in the present study as they did in Nosek (2005). To do so, we correlated the mean-level implicit attitudes obtained in the present work and in Nosek (2005) for the 54 attitude object pairs included in both studies. Despite changes in mean level implicit attitudes over time (Charlesworth & Banaji, 2019), the relationship was strongly positive and statistically significant, $r = .805$, $t(52) = 9.78$, $p < .001$, thus providing evidence for the validity of the abbreviated IATs used here.

Mixed-Effects Models

In Model 1, which used the prediction items as the only independent variable, we obtained a statistically significant and medium-sized relationship between the prediction items and IAT D scores, $\beta = .347$, $t(69.47) = 28.58$, $p < .001$. As in Studies 1A–B, based on the results of the permutation test, the significant relationship was specific to individual participants and could not be accounted for by participants predicting the general population's average IAT scores ($ps < .001$).

In Model 2, we found that the relationship between the prediction item and implicit attitudes was smaller but persisted following the addition of the explicit attitude difference scores to the model. Specifically, the effect of prediction item was reduced but remained significant, $\beta = .233$, $t(1{,}019.11) = 17.37$, $p < .001$, and explicit attitudes also had a significant effect of a similar size, $\beta = .181$, $t(9{,}053.02) = 14.53$, $p < .001$. The reduction in the coefficient for the prediction item was significant, bootstrapped 95% CI [0.093, 0.130]. Unlike in Studies 1A–1B, this pattern held regardless of whether item-level random effects were included in the model; the same applies to all further studies.

Sources of Between-Attitude Variation in Predictive Accuracy

As shown in Figure 4, predictive accuracy varied considerably as a function of the attitude object. When we regressed unstandardized D scores on unstandardized predictions for each attitude object separately, we found predictive accuracy ranging from $\beta = .010$ for the approach/avoid IAT to $\beta = .668$ for the Democrats/Republicans IAT.[9] We also refit these regression models controlling for participants' explicit attitudes, providing us, for each attitude object, a measure of participants' accuracy at predicting the portion of their implicit attitudes that diverges from their explicit attitudes (or the "uniquely implicit" component of predictive accuracy). Participants' ability to predict the uniquely implicit component of their implicit attitudes varied widely across targets as well, with the poorest performance observed for the education/defense IAT ($\beta = -.048$) and the best performance observed for the Liberals/Conservatives IAT ($\beta = .531$).

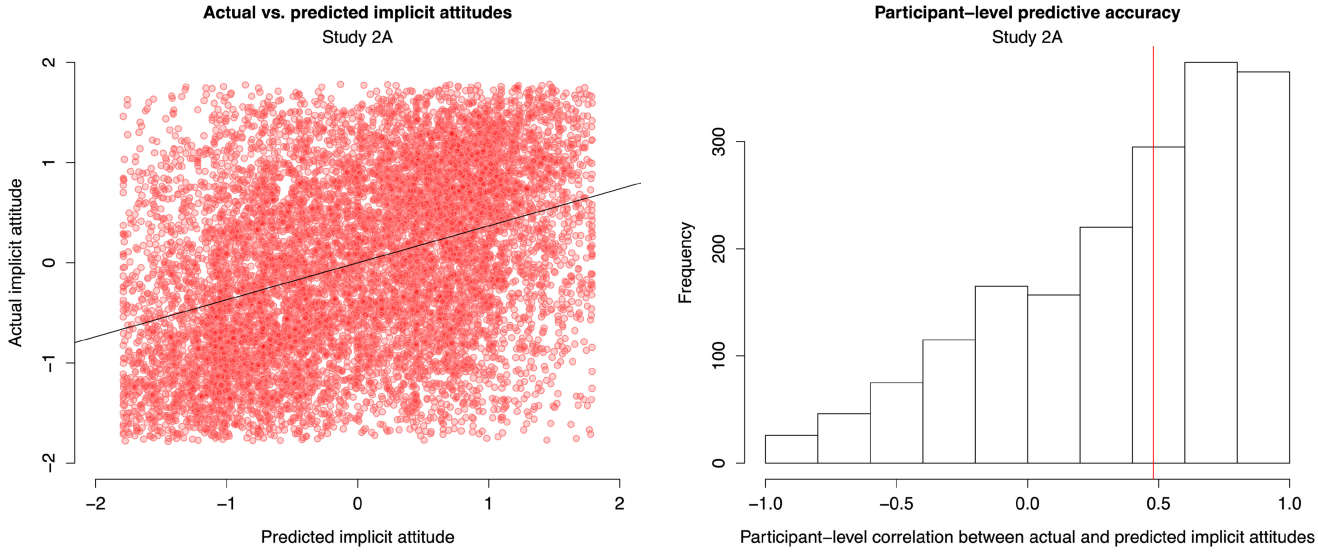
Qualitatively, participants seemed best at predicting their scores on IATs addressing political issues (including Biden vs. Trump, Liberals vs. Conservatives, creationism vs. evolution, religion vs. atheism, and feminism vs. traditional values). Participants were worst at predicting their scores on IATs indexing some social group comparisons (including rich people vs. poor people, Jews vs. Muslims, and thin people vs. fat people) and abstract comparisons that participants are unlikely to spontaneously reflect on in their daily lives (including emotions vs. reason, and public vs. private). In a set of exploratory analyses reported below, we investigated potential sources of between-attitude variability in predictive accuracy.

Demographic Predictability. We used the same analysis as in Studies 1A–B to probe the extent to which implicit attitudes were predictable based on demographic information alone. As in Studies 1A–B, unstandardized IAT scores could be predicted by demographic information, $\beta = .222$, $t(52.31) = 8.73$, $p < .001$. Importantly, due to the diversity of attitude targets in Study 2A, demographic predictability differed widely across targets, from

[9] As described above, item-level analyses require using unstandardized variables to be interpretable, as opposed to standardizing variables at the participant level (which we do for participant-level analyses; see the Method section of Study 1). Thus, it is possible that the item-level analyses reflect not only people's knowledge of their own implicit attitude scores but also their "social calibration" (i.e., knowledge of where their implicit attitude scores fall relative to the general population; Hahn & Goedderz, 2020). As discussed above, in participant-level analyses we find no difference between using standardized and unstandardized variables, suggesting that the awareness vs. social calibration distinction is not relevant to the present data.

Figure 3

Overall Correlation Between the Standardized Prediction Item and Standardized Implicit Attitudes (Left Panel) and Distribution of Participant-Level Correlations Between the Standardized Prediction Item and Standardized Implicit Attitudes (Right Panel; Study 2A)



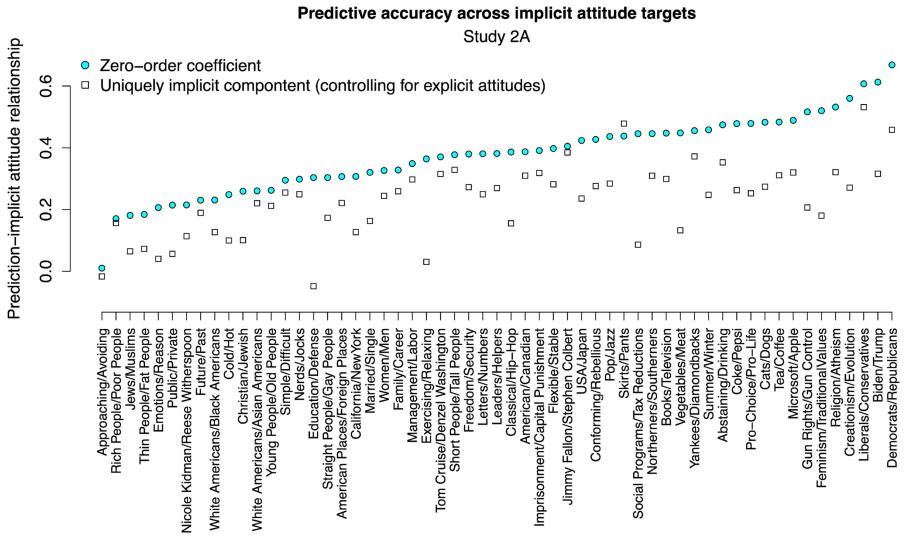
Note. See the online article for the color version of this figure.

−0.15 for the pop versus jazz comparison to 0.63 for the Democrats versus Republicans comparison. Notably, we found a significant positive correlation between the demographic predictability of a target and participants' zero-order predictive accuracy for that target (Figure 5): Participants more accurately predicted attitude targets that were in principle more predictable from demographic information, $\beta = .568$, $t(55) = 5.11$, $p < .001$.

However, this finding is complicated by two other results. First, when analyzing the uniquely implicit component of predictive accuracy (rather than zero-order accuracy), the correlation between demographic predictability and predictive accuracy dropped substantially, although it was still significant, $\beta = .358$, $t(55) = 2.84$, $p = .006$. In other words, the demographic predictability of an attitude target was strongly related to participants' ability to predict

Figure 4

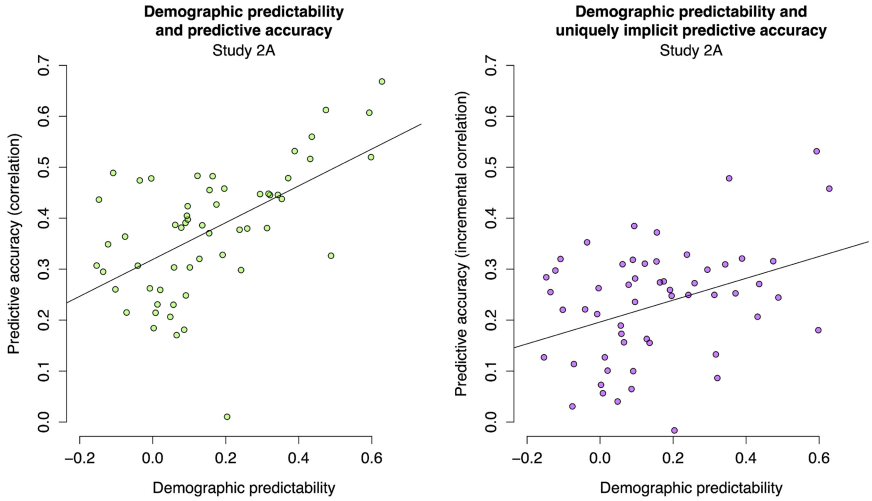
Relationship Between the Prediction Item and Implicit Attitudes at the Level of Attitude Object Pairs (Study 2A)



Note. Teal circles show the zero-order regression coefficient, and empty squares show the coefficient controlling for explicit attitudes (the "uniquely implicit" component of predictive accuracy). See the online article for the color version of this figure.

Figure 5

Correlation Between an Attitude Target's Demographic Predictability (i.e., the Correlation Between Demographic-Based Predictions and Implicit Attitudes for That Target) and Predictive Accuracy for That Target (i.e., the Correlation Between Participants' Predicted and Actual Implicit Attitudes for That Target; Study 2A)



Note. The left pane shows zero-order predictive accuracy; the right pane shows the uniquely implicit component of predictive accuracy (i.e., predictive accuracy after controlling for explicit attitudes). See the online article for the color version of this figure.

variance in their implicit attitudes that overlapped with explicit attitudes, but less so to participants' ability to predict uniquely implicit variance in their implicit attitudes (above and beyond variance accounted for by explicit attitudes).

Second, we probed whether participants accurately predicted implicit attitudes for the subset of attitude targets that could not be predicted via demographic information alone (i.e., whose coefficients were not significantly different from zero when regressing observed IAT scores on demographically predicted scores). Participants still exhibited substantial predictive accuracy, $\beta = .340$, $t(1{,}721) = 25.05$, $p < .001$. Thus, although predictive accuracy is statistically associated with the demographic predictability of comparisons, it seems that much of participants' success in predicting their implicit attitudes cannot be accounted for by demographic-based inference.

Explicit–Implicit Correlation. Next, we consider an additional variable that differs between attitude targets: the explicit–implicit correlation, or the extent to which implicit attitudes toward that target overlap with (or diverge from) explicit attitudes. We computed the explicit–implicit correlation for each of the 57 attitude targets, and then correlated this variable with participants' (unstandardized) predictive accuracy across attitude targets (see Figure 6).

The two variables were strongly positively related, $\beta = .873$, $t(55) = 13.29$, $p < .001$, indicating that participants tend to be successful at predicting the implicit attitudes that overlap most with their explicit attitudes. This relationship was strongly attenuated—but still significant—when analyzing the uniquely implicit component of people's predictive accuracy as the dependent variable, $\beta = .294$, $t(55) = 2.28$, $p = .027$. In other words, people's ability to predict variance in implicit attitudes that did not overlap with explicit attitudes was less (although still somewhat) related to those attitudes' explicit–implicit correlation.

Discussion

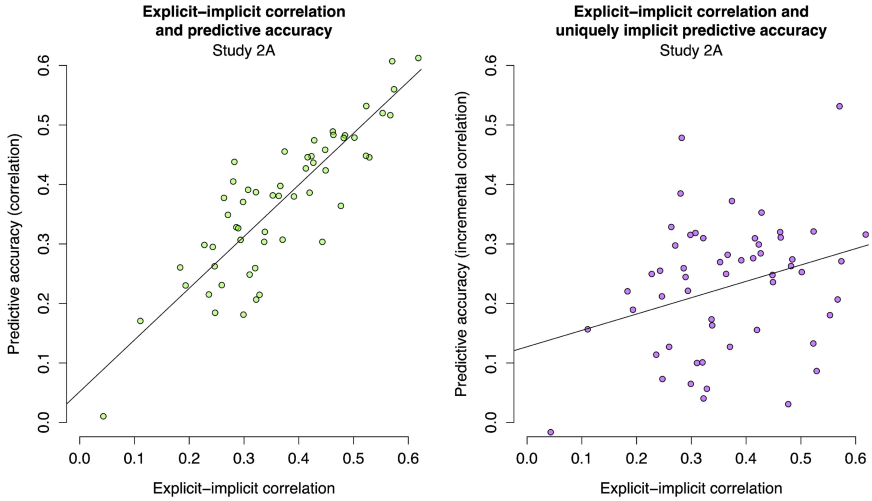
Participants were able to predict their implicit attitudes across a broad range of targets, suggesting that the results of Hahn et al. (2014) do generalize across attitude objects. In fact, Hahn and colleagues may have underestimated predictive accuracy: On average, participants were over twice as accurate in Study 2A ($\beta = .35$) than in Studies 1A–B ($\beta = .16$), which had used the limited set of targets from Hahn et al. (2014). However, there was also wide variability in predictive accuracy across attitude targets, with participants showing little accuracy for some of them. These results highlight the benefits of probing a wide range of attitude objects when investigating people's awareness of their implicit attitudes.

This attitude target-level variation in predictive accuracy allowed us to start probing the mechanisms underlying successful predictions. The exploratory results suggested that both introspective and nonintrospective mechanisms might be at play. On the one hand, predictive accuracy for each target was correlated with the target's demographic predictability, that is, the extent to which that attitude could, in principle, be predicted from demographic information alone. Moreover, predictive accuracy for each attitude target was highly correlated with the attitude's explicit–implicit correlation: Implicit attitudes with high explicit–implicit overlap (e.g., Democrats vs. Republicans) were much better predicted than those with low explicit–implicit overlap (e.g., approach vs. avoid).

This pattern of results suggests that part of participants' predictions may derive from inference over demographic information

Figure 6

Correlation Between an Attitude Target's Explicit–Implicit Correlation (i.e., the Correlation Between Explicit and Implicit Attitudes for That Target) and Predictive Accuracy for That Target (i.e., the Correlation Between Participants' Predicted and Actual Implicit Attitudes for That Target; Study 2A)



Note. The left pane shows zero-order predictive accuracy; the right pane shows the uniquely implicit component of predictive accuracy (i.e., predictive accuracy after controlling for explicit attitudes). See the online article for the color version of this figure.

and knowledge of their own explicit attitudes. This interpretation is further bolstered by the fact that participants' overall predictive accuracy was significantly reduced when controlling for explicit attitudes, suggesting that some of what people were predicting was variance in their implicit attitudes that overlapped with explicit attitudes.

On the other hand, participants exhibited substantial predictive accuracy for implicit attitudes that were plausibly difficult to infer via nonintrospective, externally observable information about the self, such as the demographic variables collected in this study (e.g., implicit preferences for emotions vs. reason). Moreover, the variables that moderated target-level variation in the zero-order relationship between participants' predictions and their IAT scores (demographic predictability and explicit–implicit correlation) had a strongly reduced effect on the uniquely implicit component of predictive accuracy.[10] This consistent and strong attenuation of the statistical relationship suggests that people's ability to predict the part of their implicit attitudes that overlap with explicit attitudes may be driven by a different mechanism from their ability to predict variance in their implicit attitudes that diverges from explicit attitudes—a mechanism that is not well captured by the predictors mentioned above.

Together, these results paint a suggestive picture of predictive accuracy. Specifically, it seems conceivable that successful predictions of implicit attitudes are driven by two separate mechanisms: (a) one that draws on more surface-level information (such as demographic information and knowledge of explicit attitudes) to predict variance in implicit attitudes that overlaps with explicit attitudes, and (b) another one that draws on direct introspective access to predict uniquely implicit variance in implicit attitudes. We continue to explore this possibility in the remainder of the studies.

Study 2B

Study 2B was conducted to probe the robustness of the findings obtained in Study 2A. Specifically, Study 2A measured participants' explicit attitudes at the end of the study, after participants had made their implicit attitude predictions (and taken the corresponding IATs). However, prior work has shown that the act of predicting implicit attitudes can bring explicit attitudes into closer alignment with implicit attitudes (Hahn & Gawronski, 2019). By measuring explicit attitudes postprediction, we may have induced an artificially high correlation between explicit and implicit attitudes and thus overestimated the degree to which explicit attitudes were involved in the process of making implicit attitude predictions. To address this concern, in Study 2B, we probed whether a measure of explicit attitudes administered prior to completing the prediction items and the IATs would produce similar results to a measure of explicit attitudes administered at the end of the study.

As an additional aim of Study 2B, we sought to directly test whether the larger effect size obtained in Study 2A (compared to Studies 1A–B) could be explained by differences in the set of attitude targets used. Since Study 2A did not include in its set of attitude targets all the original targets from Studies 1A–B, it did not allow us to directly compare participants' predictive accuracy for the five attitude targets from Studies 1A–B to their predictive accuracy for the other targets added in Study 2A. Hence, in Study 2B the set of attitude targets was expanded to also include the attitude targets from Hahn et al. (2014), resulting in a total set of 60.

[10] This was also true for moderator variables included in Nosek (2005), such as self-presentation concerns and evaluative strength. Relevant results are reported in supplemental materials.

Method

A total sample of 2,273 volunteer participants were recruited from Project Implicit (https://implicit.harvard.edu/). As preregistered, participants were excluded from analyses if they (a) did not complete all five prediction items ($n = 8$), (b) did not complete all five IATs ($n = 276$), or (c) produced response latencies of 300 ms or less on at least 10% of trials on any of the IATs, indicating careless responding ($n = 156$).

Given that some participants were excluded for multiple reasons, the final sample size was 1,827 (1,246 women, 515 men, 42 participants of other genders, and 30 participants with missing data on the gender item; $M_{age} = 34$ years, $SD = 15$). Participants were recruited without regard to country of origin, but the majority of participants ($n = 1{,}362$ or 74% of the final sample) were from the United States.

The materials, procedure, and measures were the same as in Study 2A, with the following two exceptions. First, in addition to the attitude targets used in Study 2A, three additional attitude targets were included from Hahn et al. (2014), including celebrities versus regular people, children versus adults, and White people versus Latino people, resulting in 60 total comparisons. Second, participants completed explicit measures of attitude twice: first at the beginning of the study, before reading any of the instructions ("Time-1 explicit attitudes") and second, like in Study 2A, at the end of the study, following the prediction items and the IATs ("Time-2 explicit attitudes").

Results

Given the similarity of the present findings to the findings obtained in Study 2A, the results of Study 2B are reported relatively briefly, with a focus on novel results. The details of all models are available in the online supplemental materials.

Descriptive Statistics

Like in Study 2A, the prediction item was positively correlated with IAT scores in the whole sample ($r = .337$). The median participant-level correlation was $r = .447$ ($SD = 0.474$).

Time-1 and Time-2 explicit attitudes were highly correlated with each other, $r = .807$, $t(9{,}015) = 129.77$, $p < .001$, thus suggesting that results involving the two variables will be similar. However, in line with Hahn and Gawronski (2019), we found that implicit and explicit attitudes were more closely aligned at Time-2 (following the completion of the prediction items and the IATs) than at Time-1 (at the outset of the study), $\beta = .055$, $z = 6.27$, $p < .001$. As such, comparing the two sets of explicit attitude items is of theoretical interest.

Mixed-Effects Models

In Model 1, we obtained a statistically significant and medium-sized relationship between participants' predictions and IAT scores, $\beta = .304$, $t(67.15) = 20.43$, $p < .001$. Based on the results of the permutation test, the significant relationship was again specific to individual participants and could not be accounted for by population-level associations ($ps < .001$).

In Model 2A, we found that the relationship between the prediction item and implicit attitudes persisted following the addition of the Time-2 explicit attitude difference scores to the model. Specifically, the prediction item had a reduced but significant effect, $\beta = .168$, $t(132.32) = 10.61$, $p < .001$, and explicit attitudes also had a significant effect of similar (slightly larger) size, $\beta = .205$, $t(8{,}724.61) = 15.76$, $p < .001$. The reduction in the prediction coefficient was significant, 95% CI [0.113, 0.155]. This result is highly similar to the corresponding result from Study 2A.

Critically, in Model 2B, a similar finding emerged when the Time-1 instead of the Time-2 explicit attitudes were used as controls: The prediction item had a reduced but significant effect, $\beta = .242$, $t(98.37) = 15.98$, $p < .001$, and explicit attitudes also had a significant effect (although this time the effect was smaller), $\beta = .119$, $t(8{,}739.61) = 10.00$, $p < .001$. The reduction in the prediction coefficient was again significant, 95% CI [0.046, 0.076].

In Model 2C, both sets of explicit items were considered simultaneously. In this model, the effect of the prediction item remained significant, $\beta = .167$, $t(139.50) = 10.54$, $p < .001$, and Time-2 explicit attitudes also had a significant effect, $\beta = .188$, $t(8{,}881.46) = 12.06$, $p < .001$. The effect of Time-1 explicit attitudes was not significant, $\beta = .026$, $t(8{,}892.66) = 1.84$, $p = .066$.

Subset Analyses

We did not find any evidence that the newly included attitude targets and the attitude targets originally used by Hahn et al. (2014) differed from each other in predictive accuracy, $\chi^2(2) = 0.11$, $p = .945$. As such, the difference in effect sizes between Studies 1A–1B and Study 2A cannot be explained merely by the difference in the attitude targets included.

Sources of Between-Attitude Variation in Predictive Accuracy

As in Study 2A, we found that participants were more accurate for attitude targets that were more easily predictable by demographic information, $\beta = .308$, $t(58) = 2.47$, $p = .017$, but this relationship disappeared when examining the uniquely implicit component of predictive accuracy, $\beta = -.101$, $t(58) = -0.78$, $p = .441$. Moreover, when restricting the primary participant-level analysis (i.e., Model 1 above) to only the attitude targets that could not be predicted by demographic information, people still showed substantial accuracy, $\beta = .305$, $t(10.75) = 9.15$, $p < .001$.

Of particular interest, we also found that (as in Study 2A) participants were considerably more accurate for attitude targets with a high explicit–implicit correlation, regardless of whether we used Time-1 or Time-2 explicit attitudes. When probing the attitude-level relationship between explicit–implicit correlation and zero-order predictive accuracy, the correlation was strongly positive for Time-2 explicit attitudes, $\beta = .852$, $t(58) = 12.39$, $p < .001$, and, newly, also for Time-1 explicit attitudes, $\beta = .748$, $t(58) = 8.58$, $p < .001$. However, similar to Study 2A, when analyzing the uniquely implicit component of the prediction–implicit correlation (controlling for explicit attitudes), the correlation of correlations became considerably weaker and nonsignificant, $\beta = .078$, $t(58) = 0.60$, $p = .552$, for Time-2 explicit attitudes, and $\beta = .218$, $t(58) = 1.70$, $p = .095$, for Time-1 explicit attitudes. In other words, all the results were qualitatively similar when using Time-1 versus Time-2 explicit attitudes.

Discussion

Study 2B ruled out a potential confound in Study 2A regarding the timing of explicit attitude measurement. Specifically, we found

qualitatively similar results regardless of whether explicit attitudes were measured at the beginning or end of the study. Controlling for Time-1 explicit attitudes still significantly reduced predictive accuracy (albeit to a lesser extent than Time-2 explicit attitudes did). Moreover, in the item-level analysis, predictive accuracy was associated with implicit–explicit correlations to similar degrees irrespective of whether Time-1 or Time-2 explicit attitudes were considered. Hence, the findings of Study 2A cannot be explained by the timing of explicit attitude measurement.

Study 2C

In Study 2C, we probe another element of the design of Studies 2A–B that could have influenced the results: the choice of explicit attitude measure. The relationship between implicit and explicit attitudes is known to be moderated by the type of explicit measure administered (Hofmann et al., 2005; Ranganath et al., 2008). Specifically, implicit attitudes measured by the IAT tend to be more highly correlated with feeling thermometers than with scales measuring liking; moreover, affective measures of explicit attitudes tend to be more highly correlated with IATs than cognitive measures are (Hofmann et al., 2005). Therefore, using two affective measures of explicit attitudes in Studies 1–2B—specifically, a feeling thermometer in Studies 1A–1B and a measure of liking in Studies 2A–2B—may have led us to underestimate the unique relationship of prediction items with implicit attitudes.

Therefore, in Study 2C, we randomly assigned participants to complete either the feeling thermometer measure used in Studies 1A–1B, the liking measure from Studies 2A–2B, or a newly included cognitive measure of attitudes, prompting participants to report what they think about (rather than how they feel toward) the attitude objects. As in Study 2B, explicit attitudes were measured twice—once at the outset of the study, and once following the prediction items and the IATs.

Method

A total sample of 1,721 volunteer participants were recruited from Project Implicit (https://implicit.harvard.edu/). As preregistered, participants were excluded from analyses if they (a) did not complete all five prediction items ($n = 15$), (b) did not complete all five IATs ($n = 214$), or (c) produced response latencies of 300 ms or less on at least 10% of trials on any of the IATs, indicating careless responding ($n = 132$).

Following these exclusions, the final sample size was 1,360 (953 women, 373 men, 20 participants of other genders, and 20 participants with missing data on the gender item; $M_{age} = 36$ years, $SD = 15$). Participants were recruited without regard to country of origin, but the majority of participants ($n = 976$ or 71% of the final sample) were from the United States.

The materials, procedure, and measures were the same as in Study 2B, with the following exception. Each participant was randomly assigned to complete one of three sets of explicit attitude measures: (a) the feeling thermometer measure used in Studies 1A–1B, asking participants to report how coldly or warmly they felt toward each attitude object ($n = 459$), (b) the liking measure used in Studies 2A–2B, asking participants to report how much they disliked or liked each attitude object ($n = 436$), and (c) a newly included, more cognitively oriented measure of attitudes, asking participants to report whether they believed that the attitude object was bad or good ($n = 471$). For the sake of comparability, all attitude measures were administered using 100-point sliding scales.

Results

Given the similarity of the present findings to the findings obtained in Studies 2A–2B, the results of Study 2C are reported relatively briefly, with a focus on novel results. The details of all models are available in the online supplemental materials.

Descriptive Statistics

Like in Studies 2A–2B, the prediction item was positively correlated with IAT scores in the whole sample ($r = .341$). The median participant-level correlation was $r = .440$ ($SD = 0.467$).

Time-1 and Time-2 explicit attitudes were highly correlated with each other, $r = .781$, $t(6{,}710) = 102.50$, $p < .001$; however, in line with Hahn and Gawronski (2019), we again found that implicit and explicit attitudes were more closely aligned at Time-2 (following completion of the prediction items and the IATs) than at Time-1 (at the outset of the study), $\beta = .062$, $z = 5.91$, $p < .001$.

Most relevant for the present study, we found that the three explicit attitude measures were highly correlated with each other. The attitude-level correlation between the thermometer and liking measures was 0.86 and 0.88 for Time-1 and Time-2 measures, respectively; the correlation between the thermometer and belief measures was 0.77 and 0.85 at Time-1 and Time-2; and the correlation between the liking and belief measures was 0.75 and 0.81 at Time-1 and Time-2 (all $ps < .001$). This pattern suggests that the three explicit attitude measures are unlikely to produce distinct results in the present data.

Mixed-Effects Models

Models 1 and 2 measure the relationship between participants' predictions and IAT scores, collapsing across the three explicit attitude measures. Then, Model 3 tests whether this relationship differs based on the type of explicit attitude measure.

In Model 1, we obtained a statistically significant and medium-sized relationship between participants' predictions and IAT scores, $\beta = .320$, $t(66.85) = 24.15$, $p < .001$. Based on the results of the permutation test, the significant relationship was again specific to individual participants and could not be accounted for by population-level associations ($ps < .001$).

In Model 2A, we found that the relationship between the prediction item and implicit attitudes persisted following the addition of the Time-2 explicit attitude difference scores to the model. Specifically, the effect of the prediction item was reduced but significant, $\beta = .217$, $t(138.12) = 14.45$, $p < .001$, and explicit attitudes also had a significant effect, $\beta = .176$, $t(6{,}232.09) = 12.30$, $p < .001$. The reduction of the prediction item coefficient was significant, 95% CI [0.086, 0.124].

In Model 2B, a similar finding emerged when the Time-1 instead of the Time-2 explicit attitudes were used as controls, again replicating the corresponding result from Study 2B: The effect of the prediction item was reduced but significant, $\beta = .275$, $t(553.66) = 19.29$, $p < .001$, and explicit attitudes also had a significant effect, although this time the effect was smaller, $\beta = .095$, $t(6{,}638.34) = 7.25$,

$p < .001$. The reduction of the prediction item coefficient was significant, 95% CI [0.032, 0.058].

In Model 2C, both Time-1 and Time-2 explicit items were considered simultaneously. In this model, the effect of the prediction item remained significant, $\beta = .212$, $t(143.93) = 13.92$, $p < .001$, and Time-2 explicit attitudes also had a significant effect, $\beta = .164$, $t(6{,}540.87) = 9.96$, $p < .001$. The effect of Time-1 explicit attitudes was not significant, $\beta = .023$, $t(6{,}659.10) = 1.57$, $p = .117$.

Critically, in Models 3A–3C, we tested whether any of these effects differed across the three explicit attitude measures by including interactions of the prediction item and explicit attitude item with type of explicit measure. These interactions did not significantly improve model fit over Model 2A (Time-2 explicit attitudes only), $\chi^2(6) = 4.43$, $p = .619$, over Model 2B (Time-1 explicit attitude only), $\chi^2(6) = 2.88$, $p = .824$, or over Model 2C (both explicit attitude measures), $\chi^2(8) = 3.58$, $p = .893$. Qualitatively, the results were highly similar across explicit attitude measures; for instance, the coefficient for the prediction item after controlling for Time-1 explicit attitudes was 0.27 for the feeling thermometer, also 0.27 for the liking measure, and 0.29 for the belief-based measure.

Subset Analyses

As in Study 2B, we did not find evidence that the newly included attitude targets and the attitude targets originally used by Hahn et al. (2014) differed from each other in predictive accuracy, $\chi^2(2) = 3.66$, $p = .161$.

Sources of Between-Attitude Variation in Predictive Accuracy

Importantly, we again found that participants were more accurate for attitude targets with high explicit–implicit correlations—regardless of the type of explicit attitude measure used. To test this issue, we computed the correlation between the explicit–implicit correlation and the prediction–implicit correlation at the level of attitude objects, separately for Time-1 and Time-2 explicit attitudes, first combined and then broken down by type of explicit attitude measure. When we used the zero-order relationships for each pair of variables, similar to Studies 2A–2B, the correlation of correlations was strongly positive for Time-2 explicit attitudes, $\beta = .768$, $t(58) = 9.12$, $p < .001$, and also for Time-1 explicit attitudes, $\beta = .668$, $t(58) = 6.83$, $p < .001$. However, similar to Studies 2A–2B, when analyzing the uniquely implicit component of the prediction–implicit correlation (controlling for explicit attitudes), the correlation of correlations became considerably weaker and nonsignificant, $\beta = .016$, $t(58) = 0.12$, $p = .903$, for Time-2 explicit attitudes, and $\beta = .133$, $t(58) = 1.02$, $p = .311$, for Time-1 explicit attitudes.

Critically, these results were highly similar across different explicit attitude measures. The effect of implicit–explicit correlation on predictive accuracy was uniformly significant. Using Time-2 explicit attitudes, the coefficients were $\beta = .572$, .558, and .603 for the thermometer, liking, and belief-based items, respectively; using Time-1 explicit attitudes, the coefficients were $\beta = .444$, .587, and .484. The effect of implicit–explicit correlation on the uniquely implicit portion of predictive accuracy was uniformly nonsignificant. Using Time-2 explicit attitudes, the coefficients were $\beta = .010$, $-.035$, and .177 for the thermometer, liking, and belief-based items, respectively; using Time-1 explicit attitudes, the coefficients were $\beta = .045$, .187, and .177.

Finally, unlike in Studies 2A–2B, we found that the extent to which the implicit attitudes associated with an attitude target were predictable by demographic information did not correlate with people's predictive accuracy for that target, either when examining zero-order predictive accuracy, $\beta = -.043$, $t(58) = -0.32$, $p = .747$, or the uniquely implicit component of predictive accuracy, $\beta = .010$, $t(58) = 0.075$, $p = .941$. We consider the implications of this nonreplication below.

Discussion

In conjunction with Study 2B, Study 2C underscores the robustness and generalizability of most of the findings of Study 2A. Specifically, like in Study 2B, we found that explicit attitude measures administered prior versus following the prediction items and the implicit attitude measures behaved equivalently to each other. Moreover, we were able to newly establish, contrary to some predictions derived from the relevant literature, that different types of explicit attitude measures—including a belief-based measure, a liking-based measure, and feeling thermometer measure—were also statistically indistinguishable from each other when used as control variables to determine the unique contribution of the prediction items in accounting for variance in implicit attitudes. Study 2C suggests that the key findings of Studies 2A–B concerning explicit attitudes—that is, that controlling for explicit attitudes attenuates predictive accuracy and that participants are far better at predicting their implicit attitudes for attitude items with high implicit–explicit correlations—are robust across both types of explicit attitude measures and the time of their administration.

One difference between Study 2C and Studies 2A–2B is that Study 2C did not replicate the finding that attitude targets with higher demographic predictability also produced greater predictive accuracy. This result sheds some doubt on whether people indeed use demographics to inform their predictions. To address this inconsistency, at the end of the Results section, we aggregate data from all the present studies and test whether, across all the data, item-level demographic predictability correlates with predictive accuracy. We find strong evidence that it does, suggesting that demographic-based inference plays some role in predictions. Moreover, Study 4 provides additional evidence for a small role of demographic-based inference in predictive accuracy. Nonetheless, Study 2C suggests that the item-level relationship is variable.

Study 3A

Prior work has exclusively assessed people's awareness of their implicit attitudes by comparing their predictions to the results of an IAT. However, the IAT taps more than just implicit attitudes: IAT responding emerges from multiple processes, some of which are relatively controlled, such as stimulus categorization and task switching (Conrey et al., 2005; Ito et al., 2015; Mierke & Klauer, 2001). Thus, a potential concern is that successful predictions of IAT performance are driven not by people's awareness of their implicit attitudes, but rather by some procedural feature specific to the IAT.

Rivers and Hahn (2019) provide evidence against this possibility by analyzing IAT scores with the quadruple process model and showing that people's predictions track the automatic associations component (in addition to self-control components). However, a stronger test of whether people can indeed predict the latent construct of implicit attitudes would be to probe predictive accuracy with an alternate,

independent assay of implicit attitudes. Study 3A implements this test. In Study 3A, we replicate Study 2A using an alternate measure of implicit attitudes: the AMP (Payne et al., 2005).

The AMP, a prominent alternative to the IAT for measuring individual differences in implicit attitudes, relies on the idea that people can automatically misattribute affective responses from an evaluatively meaningful stimulus to an evaluatively neutral one as long as the two are presented for short durations and in close temporal proximity to each other. For instance, if a person has an implicit attitude favoring Coke, then briefly flashing a picture of a Coke bottle will cause her to rate a subsequently presented abstract painting more favorably. Notably, the AMP shows little procedural overlap and exhibits relatively low correlations with the IAT (Bar-Anan & Vianello, 2018; Van Dessel et al., 2019), suggesting that the two measures capture relatively independent portions of the variance in implicit attitudes.

Critically, unlike the IAT, the AMP does not involve categorization of the attitude object stimuli into their respective categories (making it an irrelevant-feature paradigm; De Houwer, 2003; Field et al., 2011) or task switching. These features create a simpler procedure and can alleviate concerns about the potential contaminating effects of controlled processes specific to the IAT. Moreover, the lack of categorization of attitude object stimuli on the AMP allows us to examine whether participants are able to predict their implicit attitudes on a measure where category-level implicit attitudes emerge only to the extent that participants spontaneously categorize the stimuli in terms of the intended categories (Payne & Lundberg, 2014; Williams & Steele, 2019), thus resulting in a more conservative test. Indeed, if predictive accuracy generalizes from the IAT to the AMP, such congruence would bolster the evidence for the claim that people are aware of the underlying attitude that both measures have been designed to capture.

Method

Participants and Design

A total sample of 664 volunteer participants were recruited from Project Implicit (https://implicit.harvard.edu/). As preregistered, participants were excluded from analyses if they (a) did not complete all five prediction items ($n = 11$), (b) did not complete all five AMPs ($n = 76$), or (c) pressed the same key on all trials of at least one of the five AMPs ($n = 221$). Notably, as reported below, the results remained qualitatively the same when the latter group of participants were retained for analyses.

Following these exclusions, the final sample size was 359 (212 women, 141 men, four participants of other genders, and two participants with missing data on the gender item; $M_{age} = 39$ years, $SD = 14$). Participants were recruited without regard to country of origin, but the majority of participants ($n = 269$ or 75% of the final sample) were from the United States.

The design was the same as in Study 2A, with the exception that implicit attitudes were measured using the AMP (Payne et al., 2005) rather than the IAT.

Materials

AMP Prime Stimuli. The AMP prime stimuli were the same as the image-based category stimuli used on the IATs in Study 2A. Given that the AMP effect is hypothesized to emerge due to the misattribution of affect (Payne et al., 2005, 2013; Payne & Lundberg, 2014), the measure is more robust when the prime and target stimuli are of the same modality. As such, we omitted the word-based categories used in Study 2A from this study, resulting in 16 total comparisons included.

AMP Target Stimuli. The AMP target stimuli were 200 unique abstract paintings generated using the online service 1SecondPainting (https://1secondpainting.com/).

Procedure and Measures

The initial instructions, prediction items, explicit attitude items, and prior experience items were similar to Study 2A. However, unlike in Study 2A, the instructions and prior experience items referred to the AMP rather than the IAT. In addition, instead of the IAT, participants completed five AMPs as the main dependent measure of the study. Given that all AMPs were image-based, participants completed only one (image-based) practice prediction item.

After making their predictions, participants received the standard instructions for the AMP. That is, they were told that on each trial, they would see a painting and that their task would be to evaluate the pleasantness of that painting using the E (less pleasant than average) and I (more pleasant than average) keys. In addition, they were informed that each painting (target) would be preceded by the brief presentation of a picture (prime). Participants were asked to do their best to evaluate the targets objectively and to avoid any biasing influence of the primes.

Participants completed a total of five AMPs. On each AMP trial, a prime was presented for 75 milliseconds, followed by a blank screen for 125 ms, a target for 100 ms, and a noise mask. The noise mask remained on screen until the participant entered a response. Then the program moved on to the next trial following a 250-ms intertrial interval. Each AMP consisted of 40 trials. A randomly selected prime stimulus was presented on each trial with the constraint that across the task, each prime category was presented on 20 trials. Stimuli within each category were sampled randomly without replacement until the entire set was exhausted, at which point random selection started anew. Each target stimulus was presented once across the AMPs such that each trial featured a unique target stimulus.

An AMP score was derived by calculating the proportion of pleasant responses following each prime category and then obtaining a difference score between these two values.

Results

Descriptive Statistics

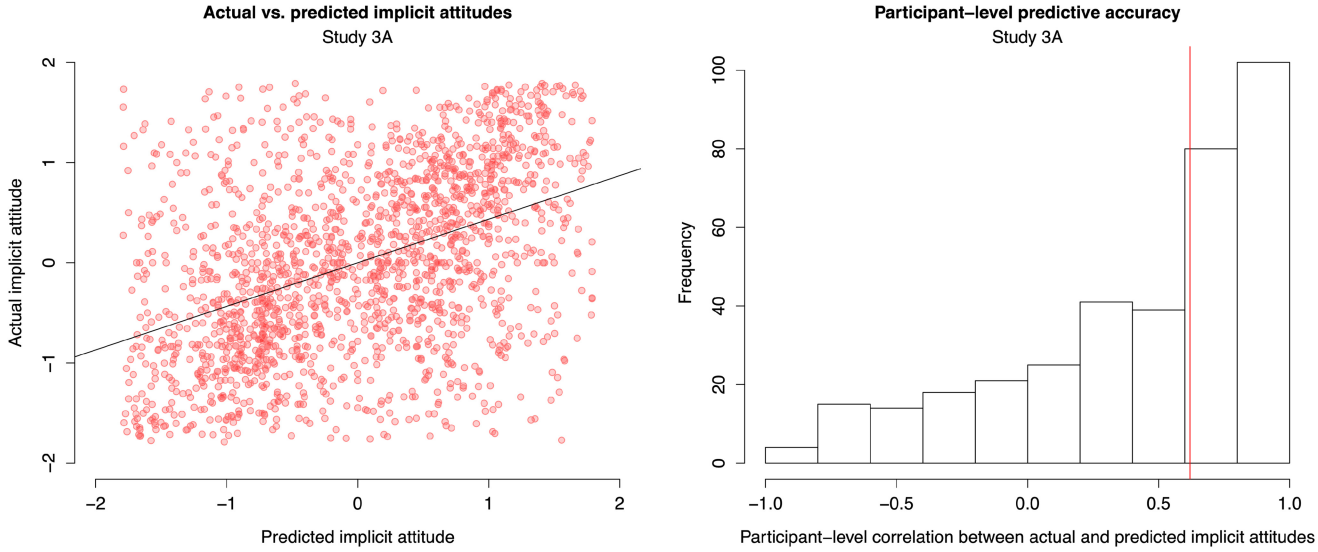
The relationship between the prediction item and implicit attitudes in the overall sample and at the participant level is shown in Figure 7. Similar to the previous studies, the prediction item was positively correlated with AMP scores in the whole sample ($r = .435$). The median participant-level correlation was $r = .619$ ($SD = 0.493$). These values are even higher than the ones obtained in Study 2A.

Mixed-Effects Models

In Model 1, we obtained a statistically significant and medium-sized relationship between participants' predictions and AMP scores, $\beta = .410$, $t(412.09) = 15.24$, $p < .001$. Based on the results

Figure 7

Overall Correlation Between the Standardized Prediction Item and Standardized Implicit Attitudes (Left Panel) and Distribution of Participant-Level Correlations Between the Standardized Prediction Item and Standardized Implicit Attitudes (Right Panel; Study 3A)



Note. See the online article for the color version of this figure.

of the permutation test, the significant relationship was again specific to individual participants and could not be accounted for by population-level associations ($ps < .001$).

In Model 2, we found that the relationship between the prediction item and implicit attitudes persisted following the addition of the explicit attitude difference scores to the model. Specifically, the effect of the prediction item was reduced but significant, $\beta = .253$, $t(705.20) = 7.89$, $p < .001$, and explicit attitudes also had a significant effect of similar size, $\beta = .246$, $t(1,757.23) = 8.67$, $p < .001$. The reduction in the prediction item coefficient was, 95% CI [0.114, 0.201].

Subgroup Analyses

We obtained the same pattern of results when analyses were restricted to (a) participants who were naïve with respect to the specific AMP, (b) fully naïve participants, (c) participants from the United States, and when (d) participants who did not have any variance in responding on at least one of the five AMPs that they had completed were included in the analyses.

Sources of Between-Attitude Variation in Predictive Accuracy

Similar to Study 2A, the size of the relationship between the prediction item and implicit attitudes varied considerably as a function of the attitude object investigated (Figure 8). Based on regression models fit to each attitude object separately, the zero-order relationship ranged from $\beta = .210$ for the thin people/fat people AMP to $\beta = .678$ for the White Americans/Asian Americans AMP.

As in Study 2A, we then refit the regression models using the uniquely implicit component of predictive accuracy (i.e., the correlation between predictions and AMP scores after controlling for explicit attitudes) as the dependent variable. We again found wide variation across attitude targets in participants' ability to predict uniquely implicit variance in implicit attitudes, with the cats/dogs AMP producing the smallest effect ($\beta = -.013$) and the Yankees/Diamondbacks AMP producing the largest effect ($\beta = .612$).

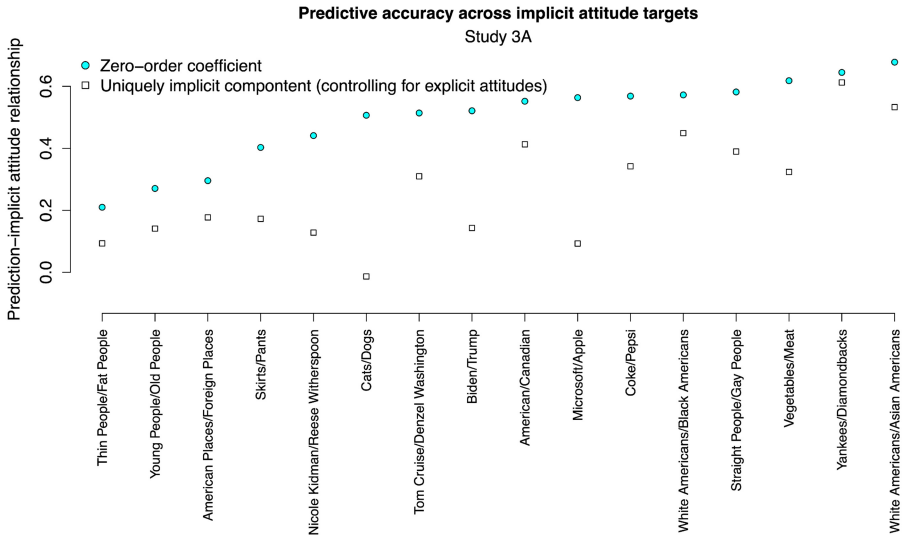
Due to the smaller number of attitude targets in Study 3A (16 vs. 57 in Study 2A), we had poor power to detect moderators of this between-attitude variation in predictive accuracy. We briefly summarize the results here for completeness, and then report them in full in the online supplemental materials. As in Studies 2A–2C, we continue to find that people were considerably more accurate at predicting their implicit attitudes for attitude targets with high explicit–implicit correlations, $\beta = .795$, $t(14) = 4.90$, $p < .001$, but this relationship disappears when examining the uniquely implicit component of predictive accuracy, $\beta = .12$, $t(14) = 0.45$, $p = .663$ (Figure 9). Moreover, we find a trend that predictions were more accurate for targets that are more demographically predictable, $\beta = .19$, $t(14) = 0.72$, $p = .483$. However, this effect is not significant, and the null finding is difficult to interpret due to very low power to detect any but the largest effects. Finally, participants still showed substantial accuracy for the subset of targets that were not predictable via demographic information, $\beta = .41$, $t(376.72) = 13.93$, $p < .001$.

Discussion

In Study 3A, participants continued to accurately predict their implicit attitudes as measured by the AMP, despite the fact that the AMP and IAT have little procedural overlap and, to a large degree, capture separate variance in underlying implicit attitudes (Bar-Anan & Vianello, 2018; Van Dessel et al., 2019). To our knowledge, this is the first demonstration that implicit attitude predictions extend beyond implicit attitudes as measured by the IAT. This result rules out the possibility that people are accurate at predicting only some specific component of the IAT, and instead bolsters the more general

Figure 8

Relationship Between the Prediction Item and Implicit Attitudes at the Level of Attitude Object Pairs (Study 3A)



Note. Teal circles show the zero-order effect (standardized regression coefficient), and empty squares show the incremental effect (standardized regression coefficient) above and beyond the effect of explicit attitudes. See the online article for the color version of this figure.

claim that people can indeed accurately anticipate their underlying implicit attitudes.

Participants again exhibited wide variation in predictive accuracy across attitude targets, and this variation showed some similar patterns as in Study 2: Specifically, predictive accuracy was better for targets characterized by high explicit–implicit correlation, with a similar (though nonsignificant) trend in demographic predictability. Given the small number of comparisons studied, these patterns should be interpreted with caution. However, with this caveat in mind, they seem largely consistent with the same picture that emerged from Study 2. Namely, predictions of implicit attitudes may be subserved by two mechanisms: one that incorporates knowledge other than direct introspection into implicit attitudes (such as explicit attitudes), and one that tracks uniquely implicit variance in implicit attitudes and involves direct introspective access to those attitudes.

Study 3B

Study 3B was analogous to Study 2B: We tested whether a measure of explicit attitudes administered prior to completing the prediction items and the AMPs would produce similar results to a measure of explicit attitudes administered at the end of the study. Moreover, like in Study 2B, the set of attitude targets was expanded to also include the five attitude targets from Studies 1A–B.

Method

A total sample of 1,091 volunteer participants were recruited from Project Implicit (https://implicit.harvard.edu/). As preregistered, participants were excluded from analyses if they (a) did not complete all five prediction items ($n = 19$), (b) did not complete all five AMPs ($n = 119$), or (c) pressed the same key on all trials of at least one of the five AMPs ($n = 426$). Notably, as reported below, the results remained qualitatively the same when the latter group of participants were retained for analyses.

Given that some participants were excluded for multiple reasons, the final sample size was 531 (333 women, 171 men, 19 participants of other genders, and eight participants with missing data on the gender item; $M_{age} = 31$ years, $SD = 14$). Participants were recruited without regard to country of origin, but the majority of participants ($n = 367$ or 69% of the final sample) were from the United States.

The materials, procedure, and measures were the same as in Study 3A, with the following two exceptions. First, in addition to the attitude targets used in Study 3A, three additional attitude targets were included from Hahn et al. (2014), including celebrities versus regular people, children versus adults, and White people versus Latino people, resulting in 19 total comparisons. Second, participants completed explicit measures of attitude twice: first at the beginning of the study, before reading any of the instructions ("Time-1 explicit attitudes") and second, like in Study 3A, at the end of the study, following the prediction items and the AMP ("Time-2 explicit attitudes").

Results

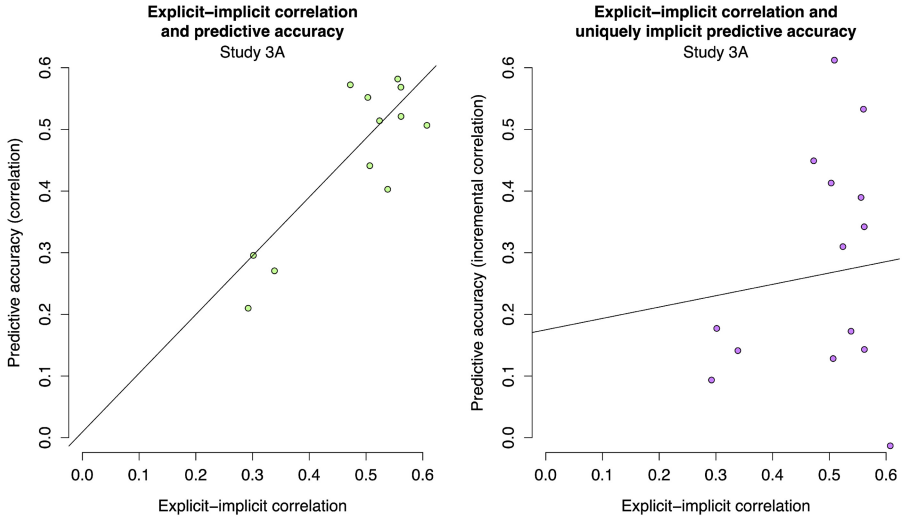
Given the similarity of the present findings to the findings obtained in Study 3A, the results of Study 3B are reported relatively briefly, with a focus on novel results. The details of all models are available in the online supplemental materials.

Descriptive Statistics

Like in Study 3A, the prediction item was positively correlated with AMP scores in the whole sample ($r = .416$). The median participant-level correlation was $r = .567$ ($SD = 0.490$).

Figure 9

Correlation Between an Attitude Target's Explicit–Implicit Correlation (i.e., the Correlation Between Explicit and Implicit Attitudes for That Target) and Predictive Accuracy for That Target (i.e., the Correlation Between Participants' Predicted and Actual Implicit Attitudes for That Target; Study 3A)



Note. The left pane shows zero-order predictive accuracy; the right pane shows the uniquely implicit component of predictive accuracy (i.e., predictive accuracy after controlling for explicit attitudes). See the online article for the color version of this figure.

Time-1 and Time-2 explicit attitudes were highly correlated with each other, $r = .827$, $t(2{,}609) = 75.42$, $p < .001$; however, in line with Hahn and Gawronski (2019), we again found that implicit and explicit attitudes were more closely aligned at Time-2 (following completion of the prediction items and the AMPs) than at Time-1 (at the outset of the study), $\beta = .069$, $z = 4.35$, $p < .001$.

Mixed-Effects Models

In Model 1, we obtained a statistically significant and medium-sized relationship between participants' predictions and AMP scores, $\beta = .379$, $t(26.67) = 12.56$, $p < .001$. Based on the results of the permutation test, the significant relationship was again specific to individual participants and could not be accounted for by population-level associations ($ps < .001$).

In Model 2A, we found that the relationship between the prediction item and implicit attitudes persisted following the addition of the Time-2 explicit attitude difference scores to the model. Specifically, the effect of the prediction item was reduced but significant, $\beta = .238$, $t(46.51) = 8.13$, $p < .001$, and explicit attitudes also had a significant effect of the same size, $\beta = .238$, $t(2{,}500.98) = 10.53$, $p < .001$. The reduction in the prediction item effect was significant, 95% CI [0.107, 0.174]. This result directly replicates the corresponding result from Study 3A.

Critically, in Model 2B, a similar finding emerged when the Time-1 instead of the Time-2 explicit attitudes were used as controls: The effect of the prediction item was reduced but significant, $\beta = .308$, $t(36.29) = 10.47$, $p < .001$, and explicit attitudes also had a significant effect, although this time the effect was smaller, $\beta = .140$, $t(2{,}551.62) = 6.73$, $p < .001$. The reduction in the prediction item effect was significant, 95% CI [0.046, 0.098].

In Model 2C, both sets of explicit items were considered simultaneously. In this model, the effect of the prediction item remained significant, $\beta = .232$, $t(47.06) = 7.87$, $p < .001$, and Time-2 explicit attitudes also had a significant effect, $\beta = .227$, $t(2{,}584.35) = 8.07$, $p < .001$. The effect of Time-1 explicit attitudes was not significant, $\beta = .022$, $t(2{,}566.94) = 0.86$, $p = .390$.

Subset Analyses

As in Studies 2B–C, we did not find any evidence that the newly included attitude targets and the attitude targets originally used by Hahn et al. (2014) differed from each other in predictive accuracy, $\chi^2(2) = 2.45$, $p = .294$.

Sources of Between-Attitude Variation in Predictive Accuracy

As in Study 3A, participants were much more accurate at predicting their implicit attitudes for attitude targets with high explicit–implicit correlations, regardless of whether Time-1 or Time-2 explicit attitudes were examined, $\beta = .767$, $t(17) = 4.93$, $p < .001$, for Time-2 attitudes, and, newly, $\beta = .725$, $t(17) = 4.34$, $p < .001$, for Time-1 attitudes. However, this relationship disappeared when examining the uniquely implicit component of predictive accuracy, $\beta = .068$, $t(17) = 0.28$, $p = .782$, for Time-2 explicit attitudes, and $\beta = -.309$, $t(17) = -1.34$, $p = .198$, for Time-1 explicit attitudes. Moreover, the correlation between attitude targets' demographic predictability and predictive accuracy was not statistically significant, $\beta = .397$, $t(17) = 1.78$, $p = .093$; however, as in Study 3A, this result is difficult to interpret due to the small sample of attitude objects. Finally, as in all prior studies, participants still showed

substantial accuracy for the subset of targets that were not predictable via demographic information, $\beta = .350$, $t(17.1) = 12.56$, $p < .001$.

Discussion

Study 3B underscores the robustness of the findings of Study 3A, just as Study 2B had for Study 2A. Specifically, when measuring explicit attitudes at the outset of the study, we replicated the findings that (a) controlling for explicit attitudes still reduced the prediction–implicit attitude relationship, and (b) across attitude items, predictive accuracy was still highly associated with implicit–explicit correlations.

Study 4

Studies 1–3 provided evidence for considerable levels of accuracy in predicting one's own implicit attitudes. Moreover, by analyzing item-level variation in predictive accuracy, Studies 2A–2C and Studies 3A–3B began to address the key question of whether accurate predictions of implicit attitudes are the result of direct introspective access to those attitudes, or whether people infer their implicit attitudes from other information about themselves (e.g., their demographics or explicit attitudes).

However, a more direct test for adjudicating between introspective and inferential mechanisms is to compare people's reports about themselves to the reports of external observers (Nisbett & Wilson, 1977). External observers can rely only on inference to predict others' implicit attitudes. Hence, to the extent that external observers can predict a person's implicit attitudes as well as a person can about herself, this pattern would demonstrate that it is possible to predict implicit attitudes without introspective access. Such a pattern of results would undermine the claim that first-person predictive accuracy is evidence for introspective awareness of implicit attitudes.[11] Moreover, to the extent that observer and first-person predictions explain overlapping variance, this would suggest that they may share a common mechanism. In contrast, to the extent that first-person predictions outperform and do not overlap with those of third-party observers, this pattern would demonstrate that first-person predictive accuracy cannot be explained by the inferential sources available to observers and would thus rule out some forms of inference as an explanation for predictive accuracy.

Following this logic, in Study 4, we ask a set of independent observers to predict other people's implicit attitudes based on information about their demographics. We test how well these observer predictions track implicit attitudes, and whether first-person predictions outperform observer predictions. We also probe the extent to which the two predictions explain overlapping or nonoverlapping variance in implicit attitudes.

Method

Participants and Design

A total sample of 1,153 volunteer participants (762 women, 375 men, 14 participants of other genders, and two participants with missing data on the gender item; $M_{age} = 40$ years, $SD = 14$) were recruited from Project Implicit (https://implicit.harvard.edu/). Participants were recruited without regard to country of origin, but the majority of participants ($n = 905$ or 78% of the sample) were from the United States.

In this study, participants predicted IAT scores not for themselves but rather for other participants selected from a previous study. For clarity, we refer to the participants in Study 4 as "observers" and the original participants with whom the observers are paired as "targets." Each observer was introduced to three target participants from Study 2A, received demographic information about them, and was asked to predict these past targets' IAT scores. Observers then completed three abbreviated IATs (one from each target) for exploratory purposes and responded to explicit items regarding their prior IAT experience. To shorten the procedure, the explicit attitude items were omitted from this study.

Procedure and Measures

Initial Instructions. The initial instructions were similar to the ones used in Studies 1B and 2, with the exception that they referenced IATs completed by the past participants rather than the participants themselves. In addition, the practice prediction items were modified to reflect the fact that the participant would make predictions for a third party rather than for herself.

Prediction Items. Observers made predictions for three randomly selected targets from Study 2A. These targets were drawn from a set of 400 participants who were randomly preselected from the Study 2A sample with the constraint that they were not allowed to have missing data on any of the demographic variables used in the description mentioned below.

For each target participant, observers were instructed to read the description and to try to imagine, to the best of their ability, what kind of person this person might be. Then they received demographic information about the target participant, including country of origin (and state of residence for U.S. targets), age, gender, race, educational attainment, occupation, political orientation, religion, and degree of religiosity. Observers were asked to visualize this person as fully as they could—what they might look like or talk like; how they might react to things; what they might believe; and what their values are. Observers were asked to spend at least 30 s to think about the target and were allowed to proceed to the next page only after 30 s had passed.

Following this screen, observers received prediction items that were modeled after the prediction items used in Studies 1B–3, with the exception that they referred to the target participant rather than the observer themselves. These items were completed in an individually randomized order and not necessarily in the order in which the past participant completed their own prediction items and the IATs. As in all previous studies, each prediction item was completed on a separate screen. Once observers completed all prediction items for one target, they moved on to the next target.

Implicit Attitudes. For exploratory purposes, observers completed three abbreviated IATs (see Studies 1A, 1B, and 2). One IAT was randomly selected from the five IATs completed by each past participant. The three IATs were completed in the same order in which observers had made predictions for the three targets.

Prior Experience. The prior experience items were the same as the ones used in Studies 1B and 2, with the exception that observers completed them for 15, rather than five, IATs.

[11] Of course, if observers could predict via inference people's implicit attitudes as well as people could for themselves, this would not alone be evidence that people actually use inference when making first-person predictions. Rather, it would demonstrate that inference is powerful enough to achieve high levels of predictive accuracy, and thus predictive accuracy is not sufficient for distinguishing between inference and introspective accounts.

Analytic Strategy

Standardization of Variables. Unlike in previous studies, where predictions scores were standardized at the participant level, in this study prediction scores were standardized at the level of each observer–target pair. This way, the analyses focus on the relative accuracy of third-party predictions made for each past participant, without taking into account how observers might use the prediction scale differently across targets. (Targets' IAT scores were still standardized for each target, as in prior studies.)

Accuracy of Observer Versus First-Person Predictions. In Model 1, targets' IAT scores were predicted in a mixed-effects model with the observers' predictions as the sole fixed effect, and an observer-prediction random slope for each participant, target, and IAT comparison. Observer-level and target-level random intercepts were omitted given standardization, but comparison-level random intercepts were included. We tested and corrected for model overparameterization in the same way as in previous studies.

In Model 2, targets' IAT scores were regressed on observers' predictions and the targets' own first-person predictions as the fixed effects; an observer prediction random slope for each participant, target, and comparison; and a first-person prediction random slope for each target and comparison. We omitted observer-level and target-level random intercepts and included comparison-level random intercepts. Using this model, we tested whether first-person predictions outperformed third-party predictions via a Wald chi-squared test for equality of coefficients in mixed-effect models (implemented in the "linearHypothesis" function of the R *car* package; Fox, 2015; Fox & Weisberg, 2018).

Overlap in Variance Explained Between Observer and First-Person Predictions. Next, to explore whether observer and first-person predictions might share a common mechanism, we tested whether they explained overlapping or nonoverlapping variance in implicit attitudes. We first probed whether including third-party predictions in the model reduces the predictive power of first-person predictions. To do so, we conducted a mediation analysis with first-person predictions as the treatment, IAT scores as the outcome variable, and third-party predictions as the mediator. Importantly, we do not use the mediation analysis to make any causal claims; rather, this analysis examines whether the unique covariance between first-person predictions and IAT scores is reduced when third-person predictions are added into the model.[12] We also reran the mediation analysis with third-party predictions as the treatment and first-person predictions as the mediator, to examine whether the unique covariance between third-party predictions and IAT scores is reduced when first-person predictions are added. The mediation analysis was conducted using the *mediate* package (Tingley et al., 2014).

Second, we computed the unique and shared variance explained by each of the two predictors using the Nakagawa and Schielzeth (2013) method implemented in the *r2glmm* package (Jaeger et al., 2017). Call the shared explained variance s, the unique first-person prediction variance f, and the unique third-party prediction variance t. We examine two amounts: $s/(s+f)$, that is, the fraction of variance explained by the first-person predictions that is shared with third-party predictions, and $s/(s+t)$, that is, the fraction of variance explained by third-party predictions that is shared with first-person predictions. We tested whether each of these quantities is significantly different from 0 and 1 by computing bootstrap CIs.

Demographic Similarity. One potential confound in these analyses is that people might know more about the attitudinal tendencies of their own demographic groups than those of other groups. For instance, a 50-year-old Asian woman might know more about the attitudinal tendencies of middle-aged Asian women than those of other demographics. If this is true, then the participants from Study 2A making first-person predictions about their own implicit attitudes would outperform the observer participants in Study 4 making third-party predictions—even if both are made purely from inference over demographic information.

To test whether people know the attitudinal tendencies of their own demographic groups better than those of other groups, we examined whether observers are more accurate at predicting the implicit attitudes of target participants who are more demographically similar to them. Specifically, we computed a measure of the demographic similarity between each observer–target pair, and regressed targets' IAT scores on the interaction between observers' third-party predictions and demographic similarity. To measure demographic similarity, we scored observer–target pairs based on whether they matched on gender, political identity, race, country of citizenship, state of residence (if in the United States), religious identity, occupation, education level, and age (within 10 years of each other). If observers' predictions are equally accurate for targets who are more or less demographically similar to them, this pattern would suggest that demographic similarity is not a concerning confound.

Results

Descriptive Statistics

The relationship between observer predictions and implicit attitudes in the overall sample and at the participant–target pair level is shown in Figure 10. The prediction item was positively correlated with IAT D scores in the whole sample ($r = .168$), although the effect was noticeably smaller than for the first-person predictions obtained in Study 2A. The median correlation at the level of participant–target pairs was $r = .231$ ($SD = 0.495$). The third-party predictions were also weakly positively correlated with targets' first-person predictions ($r = .204$).

Accuracy of Third-Party Versus First-Person Predictions

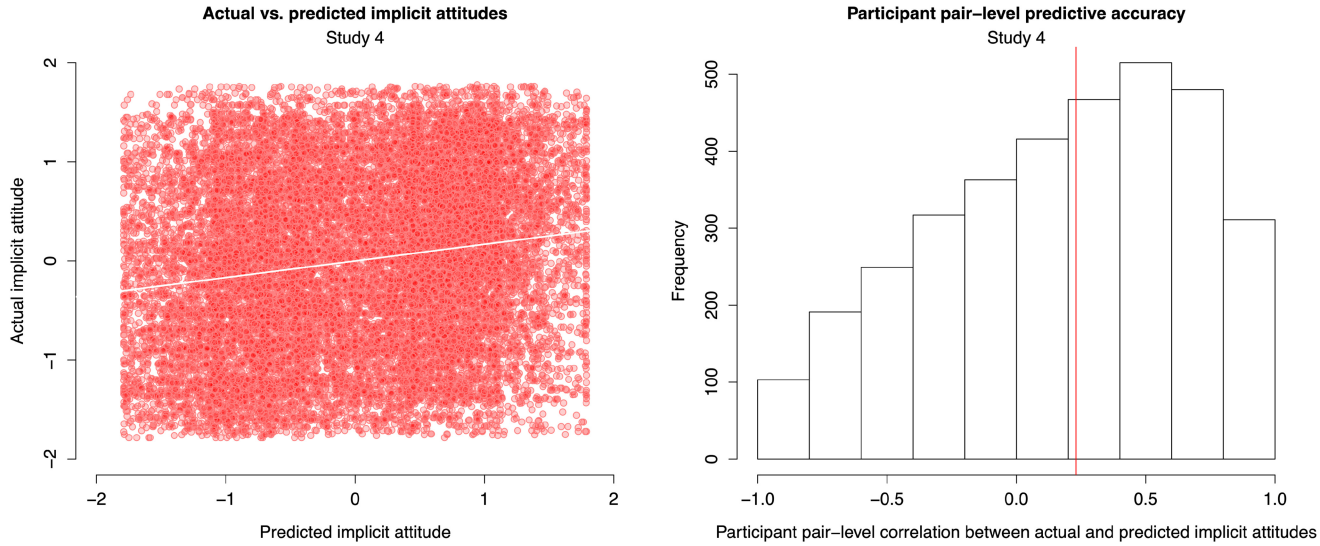
In Model 1, we obtained a statistically significant but small relationship between the third-party prediction items and IAT D scores, $\beta = .112$, $t(129.77) = 6.72$, $p < .001$.

In Model 2, we found that the relationship between the prediction item and implicit attitudes persisted following the addition of first-party prediction scores to the model. Specifically, the effect of the third-party prediction item remained significant, $\beta = .053$, $t(143.02) = 4.28$, $p < .001$. However, first-party predictions had a significant and considerably larger effect, $\beta = .358$, $t(144.41) = 11.26$, $p < .001$. The difference between the two coefficients was significant, $\chi^2(1) = 78.91$, $p < .001$.

[12] Our preregistration document indicated that we would conduct this analysis a different way; however, we realized after the fact that a proper mediation analysis would be more appropriate. The results from the preregistered analysis are similar to those reported below. (This caveat applies only to the mediation analysis; the variance decomposition analysis was preregistered.)

Figure 10

Overall Correlation Between the Standardized Third-Party Prediction Item and Standardized Implicit Attitudes (Left Panel) and Distribution of Participant-Level Correlations Between the Standardized Third-Party Prediction Item and Standardized Implicit Attitudes (Right Panel; Study 4)



Note. See the online article for the color version of this figure.

Overlap in Variance Explained Between Third-Party and First-Person Predictions

The mediation analysis found a strong direct relationship between first-person predictions and IAT scores, $\beta = .361$, $p < .001$, and a small but significant mediation effect via third-party predictions, $\beta = .014$, $p < .001$. Third-party predictions mediated about 4%, 95% CI [2%, 5%], of the relationship between first-person predictions and IAT scores.

We also refit the mediation analysis with third-party predictions as the treatment and first-person predictions as the mediator, and found that first-person predictions mediated 54%, 95% CI [45%, 65%], of the relationship between third-party predictions and IAT scores (direct effect = 0.076, $p < .001$; indirect effect = 0.089, $p < .001$). Together, these results suggest that much of third-party observers' predictions is shared with first-person predictions, and that this mechanism captures a small but nonzero fraction of first-person predictions.

The variance decomposition analysis produced a similar result. First-person predictions uniquely explained 12.2% [11.4%, 12.4%] of implicit attitude variance; third-party predictions uniquely explained 0.30% [0.19%, 0.51%]; and the two prediction types explained 1.1% [0.91%, 1.2%] shared variance. Combining these values as described in the Method section, 8.0% [7.0%, 9.4%] of the variance explained by first-person predictions was shared with third-party predictions, with the remaining 92% uniquely explained by first-person predictions. In contrast, 78.0% [70.2%, 83.5%] of the variance explained by third-party predictions was shared with first-person predictions, with the remaining 22% unique. Again, these results suggest that much of third-party observers' predictions is shared with first-person predictions, and that this mechanism underlies a small but nonzero fraction of first-person predictions.

Demographic Similarity

There was no discernible relationship between the demographic similarity between the observer–target pair and the observer's accuracy. Regressing targets' D scores on the interaction between observers' predictions and demographic similarity, the interaction was not significant, $\beta = -.003$, $t(5{,}274) = -0.66$, $p = .508$. This finding suggests that external observers were not any more successful in predicting implicit attitudes of similar than dissimilar others. As such, demographic similarity does not seem to account for the observed difference in accuracy between first-person and third-party predictions.

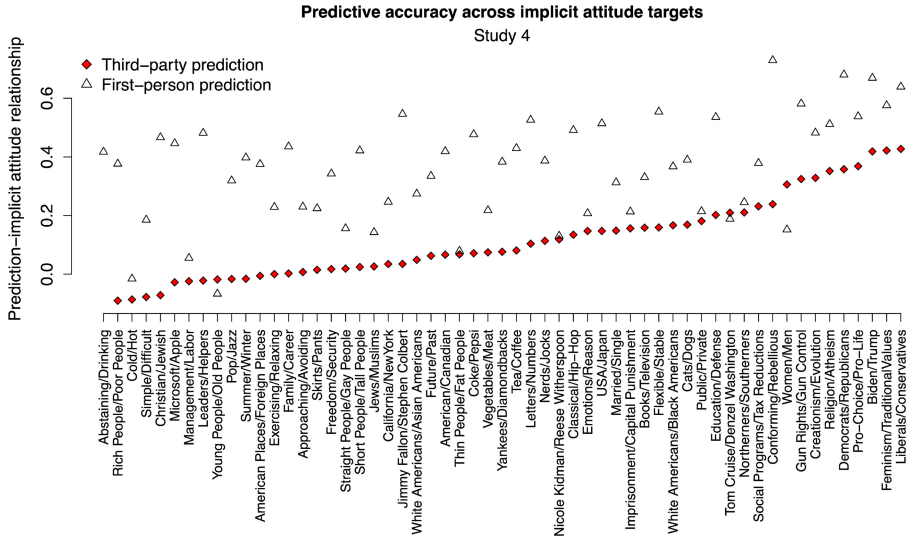
Between-Attitude Variation in Third-Party Accuracy

Based on regression models fit to each attitude object separately, the zero-order relationship between third-party predictions and IAT scores ranged from $\beta = -.147$ for the abstaining/drinking IAT to $\beta = .427$ for the Liberals/Conservatives IAT (Figure 11). Overall, similar to the first-person predictions investigated in Study 2A, the largest effects were obtained for IATs addressing political issues (including feminism vs. traditional values, Biden vs. Trump, prochoice vs. prolife, Democrats vs. Republicans, and religion vs. atheism). The smallest effects were obtained on IATs indexing some social group comparisons (including rich people vs. poor people and Christian vs. Jewish) and comparisons for which the demographic variables were arguably not particularly informative (including cold vs. hot and simple vs. difficult).

As a final test of the relationship between first-person and third-party predictions, we correlated first-person and observer predictive accuracy across attitude objects (Figure 12). When analyzing zero-order predictive accuracy, the two variables were positively correlated, $r = .490$, $t(55) = 4.17$, $p < .001$, suggesting the implicit attitudes that participants were successful at predicting for themselves also tended to be

Figure 11

Relationship Between the Prediction Items and Implicit Attitudes at the Level of Attitude Object Pairs



Note. Filled red diamonds show the third-party prediction from Study 4, and empty triangles show the first-person prediction from Study 2A. See the online article for the color version of this figure.

the ones that they were successful at predicting for third parties. This result fits with the above findings suggesting that the mechanism underlying third-party predictions is also at work in first-person predictions.

However, when we examined the uniquely implicit component of first-person predictive accuracy by partialing out explicit attitudes from the first-person prediction–implicit attitude relationship, the correlation was attenuated and became nonsignificant, $r = .218$, $t(55) = 1.66$, $p = .103$. Consistent with the results from Studies 2–3, this finding suggests that the predictive mechanism shared between observers and first-person predictors tracks the component of implicit attitudes that overlaps with explicit attitudes, but not the uniquely implicit component of implicit attitudes.

Discussion

Observers asked to predict others' implicit attitudes based on demographic information were far less accurate than people's first-person predictions about themselves, and most of the variance in implicit attitudes explained by first-person predictions was not accounted for by observers' predictions. These results provide strong evidence that people's ability to predict their implicit attitudes is not simply the result of inference over demographic information and rules out one inference-based mechanism as fully accounting for people's accuracy in predicting their implicit attitudes.

At the same time, demographic-based inference accounted for a nonzero portion of predictive accuracy. Observers' predictions did explain some of the same variance in IAT scores that first-person predictions did, and targets exhibited more predictive accuracy for attitude objects that observers could predict more accurately. These results reveal a small overlap between the processes underlying third-party and first-person predictions and suggest that not all successful implicit attitude predictions are the result of direct introspective access to those attitudes.

Overall, the results of Study 4 are again consistent with a model of implicit attitude prediction that relies on multiple mechanisms, including inference over publicly available cues (such as demographic information) and direct introspective access to implicit attitudes. Of course, people observe far more of themselves than simple demographic facts alone; therefore, the large portion of first-person predictive accuracy not explainable through demographic-based inference could still be the result of more sophisticated inferential mechanisms. We return to this point in the General Discussion section.

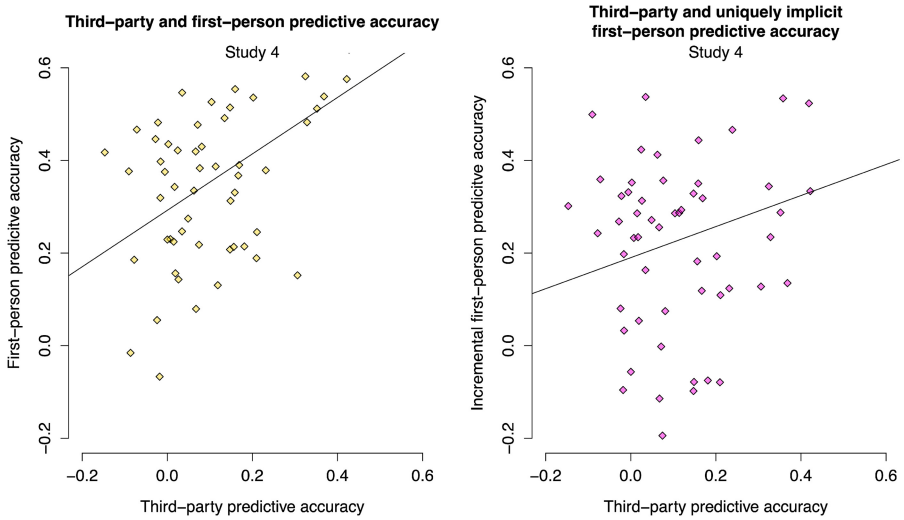
Sources of Information Underlying Implicit Attitudes Predictions: Aggregate Analysis

A final approach to characterizing the mechanisms underlying implicit attitude predictions is to regress people's predictions on the different sources of information they could have used to make those predictions and examine the percentage of variance in predictions explained by each information source. Throughout this article, we have considered three such information sources: direct introspective knowledge of implicit attitudes, explicit attitudes, and demographics. To examine the percentage of variance in predictions explained by each of these sources, we aggregated data from Studies 1–3 (seven studies, total $N = 6{,}856$) and regressed participants' implicit attitude predictions simultaneously on their actual implicit attitudes (as measured by IAT scores in Studies 1–2 and AMP scores in Study 3), reported explicit attitudes, and the predictions obtained from their demographic information (via the leave-one-out demographic analysis described in Study 1A). Details of the regression model are provided in the online supplemental materials.

All three information sources were significantly and uniquely associated with participants' predictions (for implicit attitude scores, $\beta = .177$, $t(93.56) = 20.13$, $p < .001$; for explicit attitude scores,

Figure 12

Correlation Between an Attitude Target's Third-Party Predictive Accuracy Correlation (i.e., the Correlation Between Third-Party Predictions and IAT Scores for That Target) and First-Person Predictive Accuracy (i.e., the Correlation Between First-Person Predictions and IAT Scores for That Target; Study 4)



Note. The left pane shows zero-order first-person predictive accuracy; the right pane shows the uniquely implicit component of first-person predictive accuracy (i.e., first-person predictive accuracy after controlling for explicit attitudes). IAT = Implicit Association Test. See the online article for the color version of this figure.

$\beta = .500$, $t(83.98) = 47.56$, $p < .001$; and for demographic-based predictions, $\beta = .120$, $t(60.10) = 3.84$, $p < .001$). We computed the percentage of variance uniquely explained by each regressor using the *r2glmm* package (Jaeger et al., 2017). Explicit attitudes explained by far the most variance in predictions (26.6%, 95% CI [25.7%, 27.5%]), with IAT/AMP scores explaining 4.1% [3.6%, 4.6%] and demographic-based predictions explaining 2.0% [1.0%, 3.0%].

The results were similar when restricting the analysis to Studies 2B–C and 3B, where explicit attitudes were reported prior to making implicit attitude predictions—for implicit attitude scores, $\beta = .211$, $t(159.67) = 17.48$, $p < .001$, 5.5% [4.6%, 6.6%] variance explained; for explicit attitude scores, $\beta = .448$, $t(79.52) = 27.88$, $p < .001$, 21.9% [20.3%, 23.5%] variance explained; for demographic-based predictions, $\beta = .114$, $t(57.45) = 2.83$, $p = .006$, 2.0% [1.0%, 5.0%] variance explained. Together, these findings underscore our previous conclusion that implicit attitude predictions may draw on multiple mechanisms, including both nonintrospective processes (such as inference over explicit attitudes and demographic information) and direct introspective knowledge.

Finally, we used these aggregated data to address one more question. One result that was inconsistent across the present studies was whether participants could more accurately predict their implicit attitudes for attitude targets that were more easily predictable by demographic information: Studies 2A–2B found strong evidence for this effect, but the effect did not emerge in Study 2C. The effect was also not significant in Studies 3A–B, although these null results were relatively uninformative due to poor statistical power. To produce our best estimate of this effect, we examined it in the aggregated data from Studies 1–3 and found that participants could indeed better predict attitude targets that were more demographically predictable, $\beta = .452$, $t(65) = 4.09$, $p < .001$. Although the overall effect is large, given some inconsistency across studies, it should still be interpreted with some caution.

General Discussion

Taken together, the results emerging from the present studies paint a complex picture of people's awareness of the spontaneous affective reactions comprising their implicit attitudes. On the one hand, we reveal that people's ability to predict their implicit attitudes has a substantially larger scope than uncovered in past work: People's ability to predict their implicit attitudes generalizes across diverse attitude targets, large samples, and multiple measures of implicit attitudes, including the IAT and the AMP.

Moreover, we find that this ability cannot be fully explained by the two simple inferential mechanisms we consider here: inference from demographic information and from knowledge of explicit attitudes. Specifically, participants exhibited substantial predictive accuracy for attitude targets that are difficult to predict via demographics or explicit attitudes; their first-prediction predictions explained substantial variance that did not overlap with that explained by observers' third-party predictions; and when regressing their predictions on three potential sources of information—true implicit attitudes (revealed through introspection), explicit attitudes, and demographic information—implicit attitudes explained unique variance, suggesting that predictions did not rely only on the latter two sources of information. These findings suggest that part of predictive accuracy may stem from privileged introspective access to implicit attitudes,

and are, to our knowledge, the strongest evidence to date in favor of genuine implicit attitude awareness.

On the other hand, we also find evidence that part of people's predictive accuracy is accounted by inference over explicit attitudes and demographic information. First, predictive accuracy was reduced across most of the present studies when controlling for explicit attitudes, suggesting that knowledge of explicit attitudes was informing participants' predictions. Second, people were better at predicting implicit attitudes for attitude objects that had a higher implicit–explicit correlation and that were more easily predictable via demographic information. (It should be noted, however, that these item-level analyses used unstandardized variables and thus could reflect social calibration rather than awareness [Hahn & Goedderz, 2020]; moreover, the effect of demographic predictability did not emerge in every study.) Third, the variance in implicit attitudes explained by people's first-person predictions overlapped with that explained by observers' demographic-based predictions, and people were better at predicting the attitude objects that observers could also predict. Fourth, when regressing predictions on the three potential sources of information, demographics and explicit attitudes did explain significant unique variance, with explicit attitudes explaining by far the most out of all three. If people were predicting their implicit attitudes based solely on introspection, none of these patterns would be expected.

Thus, these data are consistent with predictive accuracy emerging from both inference over demographics and explicit attitudes, and genuine introspective access to implicit attitudes. Of course, it is possible that people can additionally infer their attitudes from richer sources of information such as memories of past encounters with the attitude object or lay theories about themselves (Bem, 1972; Carruthers, 2009; Cushman, 2020; Gopnik, 1993; Nisbett & Wilson, 1977). A critical step for future work is to adjudicate between introspection and these more complex inferential mechanisms.

Additionally, the present results also reveal wide variability in people's ability to predict their implicit attitudes. For one, in close replications of Hahn et al. (2014) in Studies 1A–1B, the present effect sizes were substantially smaller than those previously reported. We discuss potential causes of this discrepancy below; regardless of the cause, future research should seek to understand the contexts that elicit lower or higher levels of predictive accuracy. Moreover, we observed wide variation in predictive accuracy across attitude targets, with some targets eliciting near-zero levels of accuracy. Finally, there was considerable variation in accuracy across participants, with many participants showing zero or negative correlations between predicted and actual implicit attitude scores. These results indicate that the contents of implicit attitudes are not uniformly accessible to introspection and highlight the need to further understand this variability.

Differences From Hahn et al. (2014) and Hahn and Gawronski (2019)

The present results differed from those of Hahn et al. (2014) and Hahn and Gawronski (2019) in several ways. First, in the close replications of Hahn et al. in Studies 1A–1B (which used nearly identical instructions and stimuli as the original experiments), the effect sizes that we observed were substantially smaller than those reported in prior work.[13] This difference could potentially be explained by the fact that the present studies were run online, whereas Hahn et al.'s were conducted in person. Online participants have been shown to be often just as attentive as in-person ones (at least on platforms involving participant payments such as Mechanical Turk or Prolific; Buhrmester et al., 2016; Crump et al., 2013; Hauser & Schwarz, 2016), and have been a reliable source of vast amounts of data on implicit social cognition (Charlesworth et al., 2023; Nosek et al., 2007), supporting the collection of larger and more diverse samples than can be accomplished on college campuses.

Nonetheless, there may be some feature of this task that limits participants' performance online (Goedderz et al., 2023). For example, perhaps introspecting on implicit attitudes requires extensive attentional resources (Goedderz & Hahn, 2022; Hahn & Goedderz, 2020; Morris, 2021), which online participants—especially the volunteers on Project Implicit—may be less willing to invest. Future work could examine this hypothesis by testing whether incentives for accuracy improve predictions. In the online supplemental materials, we enumerate all the other (minor) differences between our studies and those of Hahn et al. (2014) and discuss other potential explanations for the discrepancy in effect size.

There were two other notable discrepancies between the present results and prior work. First, in most of the present studies we found that controlling for explicit attitudes significantly reduced the relationship between predicted and actual implicit attitudes; Hahn and colleagues found that controlling for explicit attitudes did not affect this relationship.[14] Second, Hahn and Goedderz (2020) found that participant-standardized, "within-subject" analyses of implicit attitude predictive accuracy yielded lower estimates than unstandardized, "between-subject" analyses, and they interpret this as a difference between awareness (i.e., introspective, first-person knowledge of implicit attitudes) and social calibration (i.e., knowing where your implicit attitude scores fall within the studied population). In contrast, here we find no empirical distinction between the two types of analysis, and hence no empirical distinction between awareness and calibration (see the online supplemental materials for details). It is unclear which differences between the present studies and prior work explain these discrepancies; this is an important area of investigation for future work.

Understanding the Mechanisms Underlying Predictive Accuracy

When participants were able to accurately predict their implicit attitudes in the present studies, this accuracy appears to have been not purely the result of introspection. If simple inferential mechanisms (such as predicting implicit attitudes from demographics) explain part of predictive accuracy, more complex inferential mechanisms may well also be at play. A key aim for future research will be

[13] In Studies 2B–2C, we found that people could predict relative patterns of IAT scores for the five attitude objects from Hahn et al. (2014) when they were mixed in with the larger set used in the present studies, with levels of predictive accuracy more akin to those reported in Hahn et al. Nonetheless, Studies 1A–1B offer the most direct replication of Hahn et al. (2014), and thus it is worth considering potential explanations for their lower effect size.

[14] The one exception to this pattern was that, in Studies 1A–1B, controlling for explicit attitudes did not reduce predictive accuracy when omitting item-level random effects from the mixed-effects models (the analytic approach adopted by Hahn et al.). Hence, in this limited sense, the present findings did replicate those of Hahn et al. (2014). Nonetheless, the broader finding regarding explicit attitudes differed from that of Hahn et al. across most of the present studies.

to test whether people may rely on these more complex inferential mechanisms, and whether such inferential mechanisms can explain all of the observed predictive accuracy. Alternatively, as originally envisioned by Hahn et al. (2014), some or even much of predictive accuracy may be explained by direct conscious access to one's implicit attitudes.

This inquiry is complicated by the fact that the boundaries between introspecting on and inferring implicit attitudes may be fuzzy. For instance, Ranganath et al. (2008) found that people's explicit reports of their "gut" or "instant" reactions to stimuli (without any mention of implicit attitudes) correlated well with subsequent IAT scores. Do these reports count as introspections into implicit attitudes proper, or are people merely reporting content which correlates (but does not constitute) implicit attitudes? The answer depends on the correct characterization of implicit attitudes, which is contested. On the characterization we adopt here, implicit attitudes are spontaneous affective reactions (Gawronski & Bodenhausen, 2006, 2011; Hahn et al., 2014)—and hence introspecting on "gut" or "instant" reactions is synonymous with introspecting on implicit attitudes. On other characterizations, however, these gut reactions may be downstream effects of other mental events that more properly constitute implicit attitudes (e.g., the activation of associations from long-term memory; Fazio, 2007; Greenwald & Banaji, 1995), in which case knowledge of gut reactions would be a type of information from which people could infer their genuine implicit attitudes. Further unpacking the mechanisms underlying predictions will help disentangle these possibilities.

Another mechanistic question concerns the relationship between explicit attitude measures and implicit attitude predictions. Explicit attitude measures correlate strongly with predictions (Goedderz et al., 2023; Hahn et al., 2014); in the present results, we find that, when regressing predictions on implicit and explicit attitude scores, explicit attitude scores explained more than six times as much variance as implicit attitude scores did. One natural interpretation of this pattern is that, when making their predictions, people rely heavily on knowledge of their explicit attitudes, perhaps inferring their implicit attitudes in part from this knowledge.

However, alternative interpretations are also possible. First, people might be aware of their true implicit attitudes but be motivated to believe or report that their implicit attitudes are more in line with their explicit attitudes for social desirability reasons; hence, the strong influence of explicit attitudes on predictions could be distorting, not necessarily contributing to, predictive accuracy. Second, explicit attitude scores themselves may be driven in part by people's (introspection into their) implicit attitudes, which would naturally lead to shared variance between explicit attitudes and implicit attitude predictions (Greenwald & Banaji, 2017). Third, predictions may correlate more strongly with explicit than implicit attitude scores simply because they both rely on self-report, whereas implicit attitude scores are based on reaction time measures (Campbell & Fiske, 1959).[15] As such, future research should further probe the role of explicit attitudes in implicit attitude prediction.

A final open question concerns the role of attention in implicit attitude awareness (assuming, as suggested by the present data, that introspective awareness plays a nonnegligible role in predictive accuracy). One possibility is that the process of becoming aware of implicit attitudes is dependent on internal attention (Goedderz & Hahn, 2022; Hahn & Goedderz, 2020). Attention, traditionally conceptualized as a process that selects among perceptual information competing for processing (Posner, 1995), has been proposed to operate internally as well (Chun et al., 2011). In other words, there are attentional processes that select among internal representations such as mental images, task sets, and memories (Amir & Bernstein, 2021; De Brigard, 2012; Lückmann et al., 2014). In external perception, attending to a stimulus is often required to become aware of it (Dehaene et al., 2006). According to a recent proposal, the same may be true internally: There may be internal content that is unconscious due to "internal inattentional blindness" and could be brought into conscious awareness if attended to (Morris, 2021). Implicit attitudes are a prime candidate for such content.

More work needs to be done, however, to establish that implicit attitude awareness depends on attention. Hahn and colleagues have shown that people are more likely to acknowledge their biases (Hahn & Gawronski, 2019) and are less surprised about being biased (Goedderz & Hahn, 2022) after having predicted their implicit attitudes, suggesting that the prediction process had brought something "into the light" that had before been hidden to them. It is still unclear, though, whether this process involves internal attention. For one, if implicit attitude prediction involves making inferences over externally observable clues, then engaging these inferential mechanisms might be what produces the observed changes in people's acknowledgement of bias. Moreover, even if people have conscious access to their implicit attitudes, it is not clear that the introspective process is attention-dependent. Future research could test whether implicit attitude awareness correlates with attentional capacity, or whether it can be improved through attentional interventions.

Finally, even assuming that attention is involved, it is unclear exactly what people would be attending to when making their predictions. People could be attending directly to the contents of implicit attitudes (such as their spontaneous affective reactions), but they could also be attending to other content—such as explicit aspects of their attitudes, memories of past encounters with the target, etc.—that they would then use to infer their implicit attitudes. Further work is needed to disentangle these possibilities.

Between-Person Variability in Predictive Accuracy

Participants in the present studies exhibited wide variation in predictive accuracy, with participant-level correlations between predictions and attitudes ranging from highly negative to (nearly) perfect. In the present studies we did not explore this variability, focusing instead on differences in predictive accuracy across attitude objects. However, future work examining the correlates of participant-level variability could further reveal the mechanisms underlying implicit attitude predictions.

To the extent that predictions are driven by introspection, individual differences in measures of internal awareness (such as the self-reflection and insight scale; Grant et al., 2002) and motivation to direct internal attention should correlate with accuracy. Similarly, to the extent that predictions are driven by inferential mechanisms, people who are high in their ability or motivation to engage in explicit reasoning (e.g., who have a strong need for cognition; Cacioppo & Petty,

[15] In the online code (https://osf.io/gzar7/), we report an exploratory analysis of the present data suggesting that shared method variance is unlikely to fully account for the strong observed relationship between predictions and explicit attitude scores.

1982) should produce more accurate predictions. Moreover, in both cases, variation may be explained by people's different levels of understanding of the construct they are supposed to be predicting (i.e., spontaneous affective or "gut" reactions). Probing the nature of between-person variation in accuracy is an important area for future research that could bolster our mechanistic understanding of successful (and failed) implicit attitude predictions.

Limitations

As discussed above, one limitation of the present studies is their focus on demographic information and explicit attitudes as the sole sources of inferential information. These are far from the richest or most important information sources that people could use to infer their implicit attitudes. One plausible additional source is memories of past encounters with the attitude target. For example, a person could recall past instances of listening to pop and jazz music, observe that they tend to turn jazz off but turn pop up, and infer that they have an implicit preference for pop over jazz (Bem, 1972). Moreover, people have very nuanced and intricate lay theories about themselves (Argyle, 2013), and these theories could include knowledge of implicit attitudes. For instance, a White person could believe that they are the type of person who is likely to implicitly prefer pop over jazz (even if explicitly they evaluate the two equally, e.g., because of the known association between jazz and Black culture).

In practice, these information sources may be difficult to disentangle from introspection. For instance, it is possible that people's lay theories could themselves be influenced by introspective knowledge, and past encounters with attitude targets might have included introspecting on the gut reactions associated with those targets. Hence, a challenging task for future work will be to adjudicate between introspection and these more complex inferential mechanisms.

Another limitation of the present studies is that the item-level analyses—that is, testing whether people exhibit more predictive accuracy for attitude targets implicit attitudes toward which are more related to explicit attitudes or demographic variables—require the use of unstandardized variables, rather than the participant-standardized variables used in the participant-level analyses. Patterns of predictive accuracy for unstandardized D scores may reflect not only awareness of those D scores but also the social calibration needed to report where one's D scores fall relative to the general population (Hahn & Goedderz, 2020).

Concretely, for instance, it is possible that people exhibit lower accuracy for attitude targets with low explicit–implicit correlations not because they are less aware of those attitudes per se but because they are less knowledgeable about where their attitude magnitudes fall relative to the general population. However, this concern is mitigated by the fact that, in the participant-level analyses, we find no difference between using standardized and unstandardized variables, suggesting that awareness and calibration are not empirically distinct in the present data. Nonetheless, it is possible that awareness and calibration may indeed diverge in the item-level analyses, and testing this possibility is an important area for future research.

A final limitation of the present data is its correlational nature, with all the known caveats, including issues of omitted third variables and reverse causality. Specifically, we have yet to demonstrate that introspection or inferential mechanisms cause successful predictions of implicit attitudes. This limitation is, to our knowledge, shared with all existing work on implicit attitude awareness. In the next section, we suggest avenues for future research to generate causal evidence.

Suggestions for Follow-Up Work

A critical direction for future research is to further probe the mechanism(s) by which people make accurate predictions, for example, direct introspection, more complex inferential mechanisms, or a combination of both. To test for genuine introspective awareness, future work could test whether people can accurately predict novel implicit attitudes formed in the lab. People can be induced to form implicit attitudes towards novel stimuli that diverge from their explicit attitudes towards the same stimuli. For example, participants show implicit positivity toward novel targets paired with the label "privileged" (as opposed to "oppressed") while reporting explicit negativity toward the same targets (Kurdi et al., 2022; Uhlmann et al., 2006). Similarly, it has been shown that, under some conditions, diagnostic positive behavioral information can overturn negative explicit evaluations formed on the basis of facial cues while leaving implicit negativity intact (Shen & Ferguson, 2021; Shen et al., 2020). If participants were able to predict variation in their implicit attitudes toward these targets, such predictive accuracy would rule out a host of inferential mechanisms (since people have no prior experience with the targets before the experiment) and provide strong evidence for direct awareness of implicit attitudes.

Another approach could be to manipulate the instructions given to participants on how to predict their implicit attitudes. If participants instructed to introspect on their internal sensations are more accurate than participants instructed to infer their attitudes from prior experiences or theories about themselves, this would again suggest that people's predictive accuracy comes in part from introspection.

Moreover, future work should more systematically explore the moderators of predictive accuracy across existing attitude targets. None of the moderators tested here could consistently explain why participants could not predict any of the uniquely implicit component of their implicit attitudes for, for example, education versus defense, but could predict them with extraordinary accuracy for, for example, skirts versus pants. Understanding sources of this variation may provide better insight into the nature of implicit attitude prediction.

Finally, an important open question is the extent to which participants' reports of their implicit attitudes can predict their subsequent behavior in the same way that implicit measures like the IAT can (Kurdi et al., 2019). If people's reports of their implicit attitudes can predict their future behavior as well as implicit measures can (or if they can predict unique variance in future behavior controlling for scores on implicit measures), this would show that people can indeed predict a behaviorally potent component of their implicit attitudes.

Constraints on Generality

The present studies were conducted in samples of a total of over 8,000 participants recruited from 129 different countries who exhibited substantive variability on virtually all demographic variables, including gender, age, educational attainment, political

ideology, and others. As such, these samples are considerably more diverse than most experimental psychology research relying on convenience samples of college participants. At the same time, 75% of the present participants were recruited from the United States, and participants were more likely to be female than other genders, young than old, highly educated than less highly educated, and liberal than conservative. Moreover, the participants were volunteers who visited the Project Implicit demonstration website out of their own interest in learning about implicit social cognition, which may have affected their goals and motivations as they were completing the studies.

At present, especially given general convergence between the present results and results obtained in in-lab college-age samples by Hahn et al. (2014) and follow-up work, we do not see any particular reason why the results reported here would not generalize beyond the current samples. However, findings of cultural differences even in psychological phenomena previously assumed to be universal are not uncommon in experimental psychology (Henrich et al., 2010). As such, we welcome attempts to replicate and extend the present findings in diverse samples the world over and support such efforts by openly sharing data, materials, and analysis scripts.

Conclusion

The question of whether implicit attitudes are introspectively accessible is fundamental to implicit social cognition. The present results significantly bolster the evidence that people can predict their implicit attitudes, but also show that this ability is highly variable and driven, at least in part, by nonintrospective mechanisms. These findings lay the groundwork for future research into implicit attitude awareness by shifting its focus from *whether* people can accurately report their implicit attitudes to *when* and *how* they make such reports. Understanding the mechanisms underlying implicit attitude prediction is key to assessing whether people are genuinely aware of their implicit attitudes, and future research might fruitfully leverage the wide variation in predictive accuracy observed in the present studies to probe those mechanisms.

References

Alicke, M. D., & Govorun, O. (2005). The better-than-average effect. *The Self in Social Judgment*, *1*(5), 85–106.

Amir, I., & Bernstein, A. (2021). *Dynamics of internal attention and internally-directed cognition: The attention-to-thoughts (A2T) model*. PsyArXiv. https://doi.org/10.31234/osf.io/3qpgb

Argyle, M. (2013). *Lay theories: Everyday understanding of problems in the social sciences*. Elsevier.

Baars, B. J. (2005). Global workspace theory of consciousness: Toward a cognitive neuroscience of human experience. In S. Laureys (Ed.), *Progress in brain research* (Vol. 150, pp. 45–53). Elsevier. https://doi.org/10.1016/S0079-6123(05)50004-9

Banaji, M. R., Bhaskar, R., & Brownstein, M. (2015). When bias is implicit, how might we think about repairing harm? *Current Opinion in Psychology*, *6*, 183–188. https://doi.org/10.1016/j.copsyc.2015.08.017

Bar-Anan, Y., & Vianello, M. (2018). A multi-method multi-trait test of the dual-attitude perspective. *Journal of Experimental Psychology: General*, *147*(8), 1264–1272. https://doi.org/10.1037/xge0000383

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2018). *Parsimonious mixed models* [Stat]. https://arxiv.org/abs/1506.04967

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

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

Berger, J. (2020). Implicit attitudes and awareness. *Synthese*, *197*(3), 1291–1312. https://doi.org/10.1007/s11229-018-1754-3

Block, N. (1995). On a confusion about a function of consciousness. *Behavioral and Brain Sciences*, *18*(2), 227–247. https://doi.org/10.1017/S0140525X00038188

Brownstein, M. (2016). Attributionism and moral responsibility for implicit bias. *Review of Philosophy and Psychology*, *7*(4), 765–786. https://doi.org/10.1007/s13164-015-0287-7

Buhrmester, M., Kwang, T., & Gosling, S. D. (2016). *Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality data?* (pp. 133–139). American Psychological Association. https://doi.org/10.1037/14805-009

Cacioppo, J. T., & Petty, R. E. (1982). The need for cognition. *Journal of Personality and Social Psychology*, *42*(1), 116–131. https://doi.org/10.1037/0022-3514.42.1.116

Cameron, C. D., Brown-Iannuzzi, J. L., & Payne, B. K. (2012). Sequential priming measures of implicit social cognition: A meta-analysis of associations with behavior and explicit attitudes. *Personality and Social Psychology Review*, *16*(4), 330–350. https://doi.org/10.1177/1088868312440047

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*(2), 81–105. https://doi.org/10.1037/h0046016

Carruthers, P. (2009). Mindreading underlies metacognition. *Behavioral and Brain Sciences*, *32*(2), 164–182. https://doi.org/10.1017/S0140525X09000831

Charlesworth, T. E. S., & Banaji, M. R. (2019). Patterns of implicit and explicit attitudes: I. Long-term change and stability from 2007 to 2016. *Psychological Science*, *30*(2), 174–192. https://doi.org/10.1177/0956797618813087

Charlesworth, T. E. S., Navon, M., Rabinovich, Y., Lofaro, N., & Kurdi, B. (2023). The project implicit international dataset: Measuring implicit and explicit social group attitudes and stereotypes across 34 countries (2009–2019). *Behavior Research Methods*, *55*(3), 1413–1440. https://doi.org/10.3758/s13428-022-01851-2

Chun, M. M., Golomb, J. D., & Turk-Browne, N. B. (2011). A taxonomy of external and internal attention. *Annual Review of Psychology*, *62*(1), 73–101. https://doi.org/10.1146/annurev.psych.093008.100427

Conrey, F. R., Sherman, J. W., Gawronski, B., Hugenberg, K., & Groom, C. J. (2005). Separating multiple processes in implicit social cognition: The quad model of implicit task performance. *Journal of Personality and Social Psychology*, *89*(4), 469–487. https://doi.org/10.1037/0022-3514.89.4.469

Cooley, E., Payne, B. K., Loersch, C., & Lei, R. (2015). Who owns implicit attitudes? Testing a metacognitive perspective. *Personality and Social Psychology Bulletin*, *41*(1), 103–115. https://doi.org/10.1177/0146167214559712

Crump, M. J. C., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PLoS ONE*, *8*(3), Article e57410. https://doi.org/10.1371/journal.pone.0057410

Cunningham, W. A., Nezlek, J. B., & Banaji, M. R. (2004). Implicit and explicit ethnocentrism: Revisiting the ideologies of prejudice. *Personality and Social Psychology Bulletin*, *30*(10), 1332–1346. https://doi.org/10.1177/0146167204264654

Cunningham, W. A., Preacher, K. J., & Banaji, M. R. (2001). Implicit attitude measures: Consistency, stability, and convergent validity. *Psychological Science*, *12*(2), 163–170. https://doi.org/10.1111/1467-9280.00328

Cushman, F. (2020). Rationalization is rational. *Behavioral and Brain Sciences*, *43*, Article e28. https://doi.org/10.1017/S0140525X19001730

De Brigard, F. (2012). The role of attention in conscious recollection. *Frontiers in Psychology*, *3*, Article 29. https://doi.org/10.3389/fpsyg.2012.00029

Dehaene, S. (2014). *Consciousness and the brain: Deciphering how the brain codes our thoughts*. Penguin.

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

Dehaene, S., Lau, H., & Kouider, S. (2017). What is consciousness, and could machines have it? *Science*, *358*(6362), 486–492. https://doi.org/10.1126/science.aan8871

Dehaene, S., & Naccache, L. (2001). Towards a cognitive neuroscience of consciousness: Basic evidence and a workspace framework. *Cognition*, *79*(1–2), 1–37. https://doi.org/10.1016/S0010-0277(00)00123-2

De Houwer, J. (2003). A structural analysis of indirect measures of attitudes. In J. Musch & K. C. Klauer (Eds.), *The psychology of evaluation: Affective processes in cognition and emotion* (pp. 219–244). Lawrence Erlbaum Associates Publishers.

De Houwer, J., & Moors, A. (2010). Implicit measures: Similarities and differences. In B. Gawronski & B. K. Payne (Eds.), *Handbook of implicit social cognition: Measurement, theory, and applications* (pp. 176–193). The Guilford Press.

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

DiCiccio, T. J., & Efron, B. (1996). Bootstrap confidence intervals. *Statistical Science*, *11*(3), 189–228. https://doi.org/10.1214/ss/1032280214

Evans, J. S. B. T., & Stanovich, K. E. (2013). Dual-process theories of higher cognition advancing the debate. *Perspectives on Psychological Science*, *8*(3), 223–241. https://doi.org/10.1177/1745691612460685

Fazio, R. H. (2007). Attitudes as object-evaluation associations of varying strength. *Social Cognition*, *25*(5), 603–637. https://doi.org/10.1521/soco.2007.25.5.603

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

Field, M., Caren, R., Fernie, G., & De Houwer, J. (2011). Alcohol approach tendencies in heavy drinkers: Comparison of effects in a relevant stimulus-response compatibility task and an approach/avoidance Simon task. *Psychology of Addictive Behaviors*, *25*(4), 697–701. https://doi.org/10.1037/a0023285

Fleming, S. M., Dolan, R. J., & Frith, C. D. (2012). Metacognition: Computation, biology and function. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1594), 1280–1286. https://doi.org/10.1098/rstb.2012.0021

Fox, J. (2015). *Applied regression analysis and generalized linear models*. SAGE Publications.

Fox, J., & Weisberg, S. (2018). *An R companion to applied regression*. SAGE Publications.

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

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

Gawronski, B., & Bodenhausen, G. V. (2011). Chapter two—The associative–propositional evaluation model: Theory, evidence, and open questions. In J. M. Olson & M. P. Zanna (Eds.), *Advances in experimental social psychology* (Vol. 44, pp. 59–127). Academic Press. https://doi.org/10.1016/B978-0-12-385522-0.00002-0

Gawronski, B., & Hahn, A. (2018). Implicit measures: Procedures, use, and interpretation. In H. Blanton (Ed.), *Measurement in social psychology* (pp. 29–55). Routledge.

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

Goedderz, A., Azad, Z. R., & Hahn, A. (2023). *Awareness of implicit attitudes revisited: A meta-analysis on replications across samples and settings*. PsyArXiv. https://doi.org/10.31234/osf.io/frwcy

Goedderz, A., & Hahn, A. (2022). Biases left unattended: People are surprised at racial bias feedback until they pay attention to their biased reactions. *Journal of Experimental Social Psychology*, *102*, Article 104374. https://doi.org/10.1016/j.jesp.2022.104374

Gopnik, A. (1993). How we know our minds: The illusion of first-person knowledge of intentionality. *Behavioral and Brain Sciences*, *16*(1), 1–14. https://doi.org/10.1017/S0140525X00028636

Grant, A. M., Franklin, J., & Langford, P. (2002). The self-reflection and insight scale: A new measure of private self-consciousness. *Social Behavior and Personality*, *30*(8), 821–835. https://doi.org/10.2224/sbp.2002.30.8.821

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review*, *102*(1), 4–27. https://doi.org/10.1037/0033-295X.102.1.4

Greenwald, A. G., & Banaji, M. R. (2017). The implicit revolution: Reconceiving the relation between conscious and unconscious. *American Psychologist*, *72*(9), 861–871. https://doi.org/10.1037/amp0000238

Greenwald, A. G., Brendl, M., Cai, H., Cvencek, D., Dovidio, J., Friese, M., Hahn, A., Hehman, E., Hofmann, W., Hughes, S., Hussey, I., Jordan, C. H., Jost, J., Kirby, T. A., Lai, C. K., Lang, J. W. B., Lindgren, K. P., Maison, D., Ostafin, B., … Wiers, R. (2020). *The Implicit Association Test at age 20: What is known and what is not known about implicit bias*. PsyArXiv. https://doi.org/10.31234/osf.io/bf97c

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology*, *74*(6), 1464–1480. https://doi.org/10.1037/0022-3514.74.6.1464

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

Greenwald, A. G., Poehlman, T. A., Uhlmann, E. L., & Banaji, M. R. (2009). Understanding and using the Implicit Association Test: III. Meta-analysis of predictive validity. *Journal of Personality and Social Psychology*, *97*(1), 17–41. https://doi.org/10.1037/a0015575

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

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

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

Hauser, D. J., & Schwarz, N. (2016). Attentive Turkers: MTurk participants perform better on online attention checks than do subject pool participants. *Behavior Research Methods*, *48*(1), 400–407. https://doi.org/10.3758/s13428-015-0578-z

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*(2–3), 61–83. https://doi.org/10.1017/S0140525X0999152X

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

Hogg, M. A. (2016). Social identity theory. In S. McKeown, R. Haji, & N. Ferguson (Eds.), *Understanding peace and conflict through social*

identity theory: Contemporary global perspectives (pp. 3–17). Springer International Publishing. https://doi.org/10.1007/978-3-319-29869-6_1

Holroyd, J. (2012). Responsibility for implicit bias. *Journal of Social Philosophy*, *43*(3), 274–306. https://doi.org/10.1111/j.1467-9833.2012.01565.x

Holroyd, J., Scaife, R., & Stafford, T. (2017). Responsibility for implicit bias. *Philosophy Compass*, *12*(3), Article e12410. https://doi.org/10.1111/phc3.12410

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

Ito, T. A., Friedman, N. P., Bartholow, B. D., Correll, J., Loersch, C., Altamirano, L. J., & Miyake, A. (2015). Toward a comprehensive understanding of executive cognitive function in implicit racial bias. *Journal of Personality and Social Psychology*, *108*(2), 187–218. https://doi.org/10.1037/a0038557

Jaeger, B. C., Edwards, L. J., Das, K., & Sen, P. K. (2017). An R2 statistic for fixed effects in the generalized linear mixed model. *Journal of Applied Statistics*, *44*(6), 1086–1105. https://doi.org/10.1080/02664763.2016.1193725

Kropko, J., & Jeffrey, J. H. (2019). Coxed: An R package for computing duration-based quantities from the Cox proportional hazards model. *The R Journal.*, *11*(2), Article 38. https://doi.org/10.32614/RJ-2019-042

Kurdi, B., Krosch, A. R., & Ferguson, M. J. (2022). *Oppressed groups engender implicit positivity: Seven demonstrations using novel and familiar targets* [Manuscript submitted for publication]. Psychology Department, University of Illinois Urbana-Champaign.

Kurdi, B., Seitchik, A. E., Axt, J. R., Carroll, T. J., Karapetyan, A., Kaushik, N., Tomezsko, D., Greenwald, A. G., & Banaji, M. R. (2019). Relationship between the Implicit Association Test and intergroup behavior: A meta-analysis. *American Psychologist*, *74*(5), 569–586. https://doi.org/10.1037/amp0000364

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). Lmertest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1–26. https://doi.org/10.18637/jss.v082.i13

Levy, N. (2014). Consciousness, implicit attitudes and moral responsibility. *Noûs*, *48*(1), 21–40. https://doi.org/10.1111/j.1468-0068.2011.00853.x

Lückmann, H. C., Jacobs, H. I. L., & Sack, A. T. (2014). The cross-functional role of frontoparietal regions in cognition: Internal attention as the overarching mechanism. *Progress in Neurobiology*, *116*, 66–86. https://doi.org/10.1016/j.pneurobio.2014.02.002

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2017). Balancing type I error and power in linear mixed models. *Journal of Memory and Language*, *94*, 305–315. https://doi.org/10.1016/j.jml.2017.01.001

Mierke, J., & Klauer, K. C. (2001). Implicit association measurement with the IAT: Evidence for effects of executive control processes. *Experimental Psychology*, *48*(2), 107–122. https://doi.org/10.1026//0949-3946.48.2.107

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

Mooney, C. Z., Mooney, C. F., Duval, R. D., Mooney, C. L., & Duvall, R. (1993). *Bootstrapping: A nonparametric approach to statistical inference*. Sage.

Morris, A. (2021). *Invisible gorillas in the mind: Internal inattentional blindness and the prospect of introspection training*. PsyArXiv. https://doi.org/10.31234/osf.io/4nf5c

Moutoussis, M., Fearon, P., El-Deredy, W., Dolan, R. J., & Friston, K. J. (2014). Bayesian inferences about the self (and others): A review. *Consciousness and Cognition*, *25*, 67–76. https://doi.org/10.1016/j.concog.2014.01.009

Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining R2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, *4*(2), 133–142. https://doi.org/10.1111/j.2041-210x.2012.00261.x

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, *84*(3), 231–259. https://doi.org/10.1037/0033-295X.84.3.231

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

Nosek, B. A., Smyth, F. L., Hansen, J. J., Devos, T., Lindner, N. M., Ranganath, K. A., Smith, C. T., Olson, K. R., Chugh, D., Greenwald, A. G., & Banaji, M. R. (2007). Pervasiveness and correlates of implicit attitudes and stereotypes. *European Review of Social Psychology*, *18*(1), 36–88. https://doi.org/10.1080/10463280701489053

Payne, B. K., Brown-Iannuzzi, J., Burkley, M., Arbuckle, N. L., Cooley, E., Cameron, C. D., & Lundberg, K. B. (2013). Intention invention and the affect misattribution procedure: Reply to Bar-Anan and Nosek (2012). *Personality and Social Psychology Bulletin*, *39*(3), 375–386. https://doi.org/10.1177/0146167212475225

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*(3), 277–293. https://doi.org/10.1037/0022-3514.89.3.277

Payne, K., & Lundberg, K. (2014). The affect misattribution procedure: Ten years of evidence on reliability, validity, and mechanisms. *Social and Personality Psychology Compass*, *8*(12), 672–686. https://doi.org/10.1111/spc3.12148

Phelps, E. A., Cannistraci, C. J., & Cunningham, W. A. (2003). Intact performance on an indirect measure of race bias following amygdala damage. *Neuropsychologia*, *41*(2), 203–208. https://doi.org/10.1016/S0028-3932(02)00150-1

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*(5), 729–738. https://doi.org/10.1162/089892900562552

Posner, M. I. (1995). Attention in cognitive neuroscience: An overview. In M. S. Gazzaniga (Ed.), *The cognitive neurosciences* (pp. 615–624). The MIT Press.

Pronin, E., & Kugler, M. B. (2010). People believe they have more free will than others. *Proceedings of the National Academy of Sciences*, *107*(52), 22469–22474. https://doi.org/10.1073/pnas.1012046108

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

Rivers, A. M., & Hahn, A. (2019). What cognitive mechanisms do people reflect on when they predict IAT scores? *Personality and Social Psychology Bulletin*, *45*(6), 878–892. https://doi.org/10.1177/0146167218799307

Shen, X., & Ferguson, M. J. (2021). How resistant are implicit impressions of facial trustworthiness? When new evidence leads to durable updating. *Journal of Experimental Social Psychology*, *97*, Article 104219. https://doi.org/10.1016/j.jesp.2021.104219

Shen, X., Mann, T. C., & Ferguson, M. J. (2020). Beware a dishonest face?: Updating face-based implicit impressions using diagnostic behavioral information. *Journal of Experimental Social Psychology*, *86*, 103888. https://doi.org/10.1016/j.jesp.2019.103888

Tajfel, H. (2010). *Social identity and intergroup relations*. Cambridge University Press.

Tingley, D., Yamamoto, T., Hirose, K., Keele, L., & Imai, K. (2014). Mediation: R package for causal mediation analysis. *Journal of Statistical Software*, *59*(5), 1–38. https://doi.org/10.18637/jss.v059.i05

Uhlmann, E. L., Brescoll, V. L., & Paluck, E. L. (2006). Are members of low status groups perceived as bad, or badly off? Egalitarian negative

associations and automatic prejudice. *Journal of Experimental Social Psychology*, *42*(4), 491–499. https://doi.org/10.1016/j.jesp.2004.10.003

Van Dessel, P., Ye, Y., & De Houwer, J. (2019). Changing deep-rooted implicit evaluation in the blink of an eye: Negative verbal information shifts automatic liking of Gandhi. *Social Psychological and Personality Science*, *10*(2), 266–273. https://doi.org/10.1177/1948550617752064

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

Williams, A., & Steele, J. R. (2019). Examining children's implicit racial attitudes using exemplar and category-based measures. *Child Development*, *90*(3), e322–e338. https://doi.org/10.1111/cdev.12991

Wilson, T. D. (2004). *Strangers to ourselves*. Harvard University Press.

Yarkoni, T. (2022). The generalizability crisis. *Behavioral and Brain Sciences*, *45*, Article e1. https://doi.org/10.1017/S0140525X20001685

Received July 3, 2022
Revision received June 2, 2023
Accepted June 27, 2023 ■