

Self-Other Asymmetries in the Perceived Validity of the Implicit Association Test

Cristina Mendonça and André Mata
Universidade de Lisboa

Kathleen D. Vohs
University of Minnesota

The Implicit Association Test (IAT) is the most popular instrument in implicit social cognition, with some scholars and practitioners calling for its use in applied settings. Yet, little is known about how people perceive the test's validity as a measure of their true attitudes toward members of other groups. Four experiments manipulated the desirability of the IAT's result and whether that result referred to one's own attitudes or other people's. Results showed a self-other asymmetry, such that people perceived a desirable IAT result to be more valid when it applied to themselves than to others, whereas the opposite held for undesirable IAT results. A fifth experiment demonstrated that these self-other differences influence how people react to the idea of using the IAT as a personnel selection tool. Experiment 6 tested whether the self-other effect was driven by motivation or expectations, finding evidence for motivated reasoning. All told, the current findings suggest potential barriers to implementing the IAT in applied settings.

Public Significance Statement

People interpret the results of an attitudes and stereotypes measure, the Implicit Association Test, in a self-serving manner: When the test is said to reveal that one holds undesirable implicit attitudes, people think the test is less valid than when the same result is presented as referring to other people. These results suggest caution in using the test in applied settings, as people may use their self-interest in judging the test's validity.

Keywords: Implicit Association Test, motivated reasoning, self-other differences

What do people think about the use of implicit attitude tests as indicators of their and others' innermost attitudes toward certain social groups? We investigated that question using reactions to the Implicit Association Test (IAT; Greenwald, McGhee, & Schwartz, 1998), the most well-known and frequently used measure of implicit prejudice. We centered on the perceived validity of the test, meaning the extent to which people think it measures what it purports to measure. As main factors, we manipulated whether the test results reveal a desirable or undesirable level of prejudice, and whether the results refer to the self's implicit biases or other people's. We predicted that people would view the test as valid when it reveals

a desirable outcome—namely, that they are not prejudiced—whereas they would not see the test as valid when its results suggest that they hold strongly prejudicial attitudes. In altering whether the results are said to reflect one's own or others' implicit attitudes, we were able to assess whether the expected reactions would hold for the test wholesale, or whether the varying perceptions of its validity would be linked to oneself specifically. We expected the latter. That is, the IAT might be seen as invalid, for instance, when it indicates that the test taker holds prejudiced attitudes, but the same IAT result might be seen as credible when it refers to other people's attitudes.

Implicit Bias and the IAT

In 2015, the White House Office of Science and Technology (Handelsman & Sakraney, 2015) produced a report on implicit bias, explaining the concept, how to measure it, its impact, and how to reduce it. This is a rather impressive development, given that contemporary research on implicit social cognition, the field that gave rise to the notion of implicit bias, was launched a mere two decades prior (Greenwald & Banaji, 1995). The White House was not alone: Google, Facebook, Microsoft, Coca-Cola, and Proctor & Gamble are examples of organizations that have offered or mandated training to their employees with the aim of reducing implicit bias (Staats, Capatosto,

Cristina Mendonça and André Mata, CICPSI, Faculdade de Psicologia, Universidade de Lisboa; Kathleen D. Vohs, Carlson School of Management, University of Minnesota.

Cristina Mendonça and André Mata contributed equally to this work.

This research was supported by the following grants IF/01612/2014 and PD/BD/113491/2015 from the Portuguese Science Foundation.

Correspondence concerning this article should be addressed to André Mata, CICPSI, Faculdade de Psicologia, Universidade de Lisboa, Alameda da Universidade, 1649-013 Lisboa, Portugal. E-mail: andremata@psicologia.ulisboa.pt

Wright, & Jackson, 2016). This rise in societal and organizational interest may be tied back to scholarly claims that implicit bias is pervasive. For instance, a study of thousands of people found that 80% of respondents showed a negative implicit attitude toward elderly people and that more than two thirds (68%) hold a negative implicit attitude toward Blacks (Nosek et al., 2007). Research has also shown that implicit biases can influence judgment and behavior in the absence of consciousness or intentional control (Greenwald & Krieger, 2006) and can evince concrete consequences such as nonverbal hostility (Dovidio, Kawakami, & Gaertner, 2002) and curtailed job opportunities (Rooth, 2007).

Most research on implicit bias uses an implicit social cognition measure, the IAT (Greenwald et al., 1998). Generally speaking, people use keys on a computer keyboard to categorize stimuli into four categories. Two categories represent target groups (e.g., Whites and Blacks) and the other two represent attributes (e.g., pleasant and unpleasant). In the critical blocks, the same computer key corresponds to one target and one attribute category. In half of the critical blocks, the stimuli and categories are paired in a manner that is congruent with a given stereotype or biased attitude (e.g., pleasant and White are assigned to the same key, and unpleasant and Black to another key). In the other half, they are paired in a way that is incongruent with the stereotype or biased attitude (e.g., pleasant is paired with Black, and unpleasant with White). If people are faster in congruent trials than incongruent trials, it suggests a difference in the strength of association between the stimuli and mental categories, which is interpreted as reflecting an attitude (Greenwald et al., 2002). Keeping to the White–Black example, faster reaction times (RTs) for congruent compared to incongruent blocks (e.g., when using the same key to respond to ethnically Black words and to unpleasant words yield faster RTs than Black-related words and pleasant words) are seen as reflecting a more positive implicit attitude toward Whites than Blacks. Stereotypes are measured by replacing the affective attribute categories (e.g., pleasant and unpleasant) with attribute categories that describe the content of stereotypes (e.g., hostile and calm).

Notwithstanding criticisms as to whether the IAT does in fact reflect implicit attitudes (Blanton, Jaccard, Gonzales, & Christie, 2006; Brendl, Markman, & Messner, 2001; Fiedler, Messner, & Bluemke, 2006), its implications are taken seriously in academia and applied settings. Scholars and practitioners are considering interventions to correct or safeguard against implicit biases in arenas as diverse as the media (Kang, 2005), law (Kang et al., 2012; Norton, Sommers, & Brauner, 2007; Rachlinski, Johnson, Wistrich, & Guthrie, 2009), and hiring policies (Kang & Banaji, 2006; Krieger & Fiske, 2006). Moreover, scholars have recommended using the IAT as an intervention tool to raise consciousness about implicit bias (Hillard, Ryan, & Gervais, 2013; Morris & Ashburn-Nardo, 2010; Saujani, 2003), and even to prevent biased individuals from occupying positions where bias may result in significant consequences (Ayres, 2001; Saujani, 2003).

The research presented in this article tested the perceived validity of the IAT, that is, the extent to which the IAT is seen as revealing something meaningful about how people feel about members of certain groups, or whether it is discredited and

rejected. We posited that the desirability and source of an IAT result (oneself or others) would have a significant impact on the IAT's perceived validity.

Perceptions of Validity

The key outcome in this work is laypeople's validity perceptions of the IAT after information about the test's results is provided. By perceived validity, then, we mean the extent to which people believe that the IAT measures what it purports to measure (e.g., attitudes). These perceptions of validity may bear on persuasion, attitude change, and behavior following a message that involves the IAT or its results. For example, messages seen as credible can affect trust in the source and adoption of suggestions that follow from the message (see Pornpitakpan, 2004, for a review). In another example, perceived validity of a personnel selection process has been found to have an impact on organizational attractiveness, willingness to recommend the employers to others, and procedural and distributive justice perceptions (Smither, Reilly, Millsap, Pearlman, & Stoffey, 1993). Thus, perceived validity of the IAT and the factors that influence it may have important consequences for attempts to use the IAT in applied settings.

A host of factors predict whether messages and their sources will be seen as valid and credible. Germane to the current work are findings on messages with high personal relevance, which people tend to process deeply and systematically (e.g., Petty & Cacioppo, 1979). Deep cognitive processing, however, does not necessarily lead to an unbiased appreciation or understanding of the message. Rather, when people have a vested interest in an outcome, such as when it is said to reveal an aspect of their personality, people likely will engage in biased processing (Kunda, 1990). A message containing self-threatening feedback often elicits defensiveness, resulting in it being ignored or discredited as a way to lessen the implications for the self-concept (Baumeister, Bratslavsky, Finkenauer, & Vohs, 2001). By contrast, feedback that reflects positively on the self tends to be accepted without much skepticism (Kunda, 1990).

Desirability

One of the main advantages of implicit measures is that they may uncover attitudes and stereotypes that people may not wish to have (Greenwald & Banaji, 1995). Moreover, people may view evidence of some implicit associations as undesirable because they suggest attitudes that clash with their explicit attitudes and values, especially for biases that are socially condemnable (Crandall, Eshleman, & O'Brien, 2002; Zitek & Hebl, 2007). IAT scores may reveal implicit associations that predict unfavorable hiring decisions (e.g., Bendick & Nunes, 2012; Ziegert & Hanges, 2005), differential evaluation of legal evidence by jurors (e.g., Morrison, DeVaul-Fetters, & Gawronski, 2016; Rachlinski et al., 2009), and low-quality health care for obese people (Teachman & Brownell, 2001), as well as other phenomena, such as voting behavior (Friese, Bluemke, & Wänke, 2007; Friese, Smith, Plischke, Bluemke, & Nosek, 2012) and consumer choices under cognitive load (Gibson, 2008).

We expected that the desirability of the results would be a prime predictor of how valid people find the IAT. Specifically,

if the IAT reveals an undesirable association (e.g., that the test-taker has a negative attitude toward Blacks), people should perceive it as an invalid measure of their attitudes. By contrast, if the IAT is said to reveal a desirable pattern (e.g., that an egalitarian White test-taker has a positive attitude toward Blacks), we expected people would find the test's validity to be higher. This hypothesis is grounded in theories of motivated reasoning, according to which people interpret personally relevant information in self-serving ways and reject information with unfavorable self-relevant implications (e.g., Ditto & Lopez, 1992; Kunda, 1990; Lord, Ross, & Lepper, 1979; Mata, Ferreira, & Sherman, 2013; Mata, Sherman, Ferreira, & Mendonça, 2015). Indeed, people react in a self-serving manner to cognitive and personality tests used in job applicant testing as a function of their expected or obtained results (Chan, 1997; Chan, Schmitt, DeShon, Clause, & Delbridge, 1997; Chan, Schmitt, Jennings, Clause, & Delbridge, 1998). Thus, we expected judgments of the IAT's validity to vary as a function of the desirability of the test results. For example, if the IAT reveals an undesirable association (e.g., that the test-taker has a negative attitude toward Blacks), people should perceive it as an invalid measure of their attitudes.

Self-Other Differences

Although desirability is a potential factor for whether people accept the conclusions suggested by their IAT score, that may not be the case when the score refers to others' scores. People may reject an IAT result if it suggests negative things about themselves, whereas the same IAT score said to reflect others' attitudes may be seen as perfectly acceptable. Therefore, people might consider the IAT to be valid when it suggests that others are biased. There is abundant evidence of self-other differences in reactions to esteem-threatening information showing that people think in strategic ways to deflect negative conclusions for themselves, but not for others (e.g., Mata, Simão, Farias, & Steimer, 2018; Steimer & Mata, 2016).

In addition, there are a host of comparative biases whereby people see others as inferior and biased. For example, people believe that others adhere less to social norms (Codol, 1975) and are more susceptible to reasoning and judgment biases (Mata, Fiedler, Ferreira, & Almeida, 2013; Pronin, Lin, & Ross, 2002) compared to oneself. Particularly relevant to the present research, people suspect that others harbor more prejudice than they do. People expect others' attitudes to be more pro-White/anti-Black than their own (Nosek & Hansen, 2008), and to exhibit more pro-White preference on the IAT than they themselves would (Hahn, Judd, Hirsh, & Blair, 2014). An IAT score suggesting that another person is biased would confirm these expectations, and is therefore likely to be seen as credible.

When it comes to self-perceptions, respondents can become defensive when confronted with an IAT score suggesting that they are biased (Hillard et al., 2013; Howell, Gaither, & Ratliff, 2015; Perry, Murphy, & Dovidio, 2015; see also Howell et al., 2013). For instance, the larger the discrepancy between explicit and implicit attitudes in a pro-White direction, the greater people's defensiveness about the test (Howell et al., 2015). This defensive response can be attenuated if the IAT result is framed as resulting from factors outside the self (e.g., cultural knowl-

edge) instead of resulting, exclusively, from personal beliefs (Hillard et al., 2013). Thus, we predicted that the perceived validity of the IAT would vary depending on whom it refers to, such that the same IAT result may be credible for the self but not credible for others, or vice versa.

Potential Individual-Difference Predictors of Perceived Validity

Although the focus of this article is on the role of desirability and source of IAT results, there are individual-difference variables that may also influence whether people accept a certain IAT score. The present studies tested their potential effect and used them as statistical controls when testing the main hypotheses.

One potentially relevant individual difference is people's explicit attitudes. People might consider an IAT result that aligns with their espoused attitudes to be more valid than one that does not. This hypothesis is consistent with findings showing that people who received IAT feedback closer to their explicitly reported attitudes showed less derogation of the IAT than those with a larger mismatch (Howell et al., 2015).

Implicit attitudes may also influence how people rate the IAT's validity, especially if people are aware of their implicit bias. Of the few studies that have investigated whether people are aware of their implicit biases, some suggest that taking the IAT can heighten awareness of one's bias. People reported feeling greater difficulty in trials congruent with stereotypes than in trials with stereotype incongruent pairings (Monteith, Voils, & Ashburn-Nardo, 2001). Moreover, there is evidence that people can predict their IAT score without even having any previous experience with the instrument and simply upon reading a minimal description of the task (Hahn et al., 2014).

Still, evidence for the awareness of one's implicit bias is not consistent. For instance, Howell et al. (2013) found that participants' IAT result predictions were uncorrelated with their actual results and Perry et al. (2015) did not find a significant correlation between IAT scores and a self-awareness scale about subtle biases against Blacks. Furthermore, even though many participants in Monteith et al.'s (2001) study detected that they performed better in congruent than in incongruent blocks, most did not interpret their IAT score as resulting from racial or stereotypic associations. Because of this mixed evidence as to whether, or to what extent, people are aware of their implicit biases, we explored both the direct impact of implicit attitudes and its indirect impact through implicit attitude awareness on perceived validity of the IAT.

An additional individual factor that we considered is the extent to which people are driven by an internal or an external motivation to respond without prejudice (Plant & Devine, 1998). Society conveys norms regarding proscribed or prescribed targets of negative attitudes (Crandall et al., 2002). Although some people adhere to these norms because they have internalized them, others adhere to them due to social pressure (Plant & Devine, 1998). These motivations are associated with different levels of implicit and explicit bias (Devine, Plant, Amodio, Harmon-Jones, & Vance, 2002), with whether people report similar opinions in private versus public contexts (Plant & Devine, 1998), and with what people perceive to be prejudiced behavior (Mitamura, Erickson, & Devine, 2017). Because these motivations may lead to

different reactions to the IAT scores depending on the situation (e.g., lower IAT derogation of a socially undesirable IAT result received in private for individuals high in external motivation), we attempted to explore their impact on perceived validity.

Research Aims and Potential Advancements

The present research aimed to add to the literature in several key ways. First, it tested whether the perceived validity of the IAT varies as a function of whether the results are about oneself versus others. We predicted that although people may reject an undesirable IAT result suggesting that they are biased, they might regard that same result as credible when it refers to others. If so, this may have societal consequences: For instance, an IAT score suggesting bias might be deemed as invalid by someone to whom the score applies, for instance a supervisor or legal defendant, but the same IAT score might be seen as compelling evidence by a company's human resources division, or plaintiffs, prosecutors, members of the discriminated group, judges, or juries. To our knowledge, there is no prior research that investigated self-other differences in people's reactions to the IAT.

Second, this research examined the role of motivation by manipulating the desirability of an IAT result to test whether people accept IAT results when they are desirable but reject them when they are undesirable. Most studies on how people react to the IAT (Hillard et al., 2013; Howell et al., 2013, 2015) did not manipulate the desirability of the test results (for an exception, see Perry et al., 2015, Study 2). The present studies introduced a novel test of the role of motivation by manipulating desirability independently from the positivity or negativity of the IAT result (Experiment 4) and expectations of one's own or others' attitudes (Experiment 6).

Third, we assessed individual differences that could potentially relate to people's perceptions of the IAT's validity and controlled for their effect when testing the predicted self-other asymmetry. To make sense of a given IAT result, people may first use their explicitly held attitudes, judging the IAT's validity as a function of the test results' alignment with their explicit attitudes (Experiments 1, 3, and 4). People may also use their implicit attitude, either directly (tested in Experiments 1–6) or by consulting their (introspected) expectation of their implicit attitude (Experiments 3–4). Experiment 4 also assessed how internal and external motivations to respond without prejudice influence the way people react to the IAT. Although some studies have explored the role of individual differences (e.g., Howell et al., 2015), none had explored such a wide set of variables known to be important for stereotyping and prejudice.

Fourth, we tested whether the impact of desirability and source (self or others) of IAT result on perceived validity would have downstream effects on acceptance of the IAT in applied settings. Experiment 5 investigated whether perceiving the IAT as a valid measure of attitudes alters beliefs about its suitability as a tool to select workers for jobs where bias might be a factor. We expected that people's perceptions of the IAT's perceived validity would carry over to their opinion about whether it should be used as a personnel selection instrument.

Last, this work revisited a critical debate pertaining to whether these effects are indeed motivational in nature or rather can be explained by beliefs and expectations (Chambers & Windschitl, 2004; Ditto, Munro, Apanovitch, Scepansky, & Lockhart, 2003; Erdelyi, 1974; Miller & Ross, 1975; Nisbett & Ross, 1980; Tetlock

& Levi, 1982). Experiment 6 gave people IAT results that were either more positive or negative than what they expected their attitudes, or others', to be. To the extent that desirability has an effect over and above expectations in how participants react to IAT results, people should be more accepting of desirable results than of undesirable ones, even if both are equally inconsistent with their expectations. This approach represents a departure from previous studies by directly testing motivation independent of belief as a determinant of reactions to the IAT.

Experiment 1

Experiment 1 had people complete an implicit measure (the IAT) as well as an explicit measure of Black–White racial attitudes. People assigned to the self condition then received alleged feedback about their scores on both measures, whereas people assigned to the others condition received an IAT result said to represent other people's typical scores. In both conditions, the bogus results mimicked the most common findings in the literature, namely low levels of prejudice on the explicit measure and high levels of anti-Black bias on the IAT. The key outcomes were perceptions of each instrument as a valid measure of prejudice. Our hypothesis was that a self-other asymmetry should emerge, such that people would find the IAT to be less valid of their true attitudes than of others'.

Two ancillary predictions also were tested. We expected that people's explicit attitudes would influence their perception of the IAT's validity, such that those with more positive explicit attitudes would consider a negative implicit attitude result as less valid. Implicit attitudes may also independently influence perceived validity in the same direction as explicit attitudes, especially if people are, in general, aware of them.

Method

Participants. Sixty participants (24 women, 36 men, $M_{\text{age}} = 30.75$, $SD_{\text{age}} = 8.59$, age range: 19–57 years) were recruited through Amazon's Mechanical Turk (MTurk). Sample size was not determined by a power analysis, but rather by a simple a priori rule of 30 participants per condition. Participants were paid \$1.33 for an estimated duration of 20 min (real average 12 min). Participants were required to (a) be located in the United States, (b) have an approval rate greater than or equal to 95%, and (c) have satisfactorily completed more than 100 prior tasks, in line with usual recommendations (e.g., Peer, Vosgerau, & Acquisti, 2014). Forty-one participants identified as White or Caucasian, seven as Black or African American, five as Hispanic, five as Asian, and two as mixed race.

Eight participants were removed from analyses for having RTs under 300 ms in 10% or more of the critical trials of the IAT, as advised by Greenwald, Nosek, and Banaji (2003).¹ An extra seven

¹ Re-running the regression analyses without any exclusion criteria, throughout our experiments, did not change any statistical inferences regarding our experimental conditions and the key interaction. As for the *t*-tests, the results are overall the same, but in Experiment 4 the significant self-other difference in the Blacks/negative condition becomes nonsignificant, $p = .055$, whereas the nonsignificant self-other difference in the racists/positive condition becomes significant, $p = .036$, and in Experiment 6 the nonsignificant effect in the desirable deviation becomes significant, in the expected direction, $p = .008$.

participants started the experiment but quit before the point where the procedure differed between conditions, leading to a low possibility of selective attrition (Zhou & Fishbach, 2016). After exclusions, our main analyses had 80% power to detect effect sizes with $f^2 = .28$ (between medium, .15, and large, .35; for a regression with 5 parameters and 52 participants), which should be adequate to detect the typically large self-other asymmetries (see Heine & Hamamura, 2007, and Study 3 of Hahn et al., 2014), but may not be enough to detect the effect of the individual-differences variables (a meta-analysis of individual differences is provided before the General Discussion to deal with this issue).

Materials and procedure. Participants started by completing a race IAT using the procedure and materials in Greenwald et al. (1998),² without trial feedback (all IAT scores in all our studies, unless otherwise stated, were calculated using D_4 to account for this aspect), described as a simple word categorization task. They completed seven blocks: (a) a 20-trial block in which they categorized names as to whether they belonged to Black (e.g., “THEO”) or White racial groups (e.g., “JONATHAN”); (b) a 20-trial block in which participants categorized words, written in lowercase, as to whether they belonged to pleasant (e.g., “paradise”) or unpleasant categories (e.g., “rotten”); (c) a 20-trial block in which all stimuli were presented and the categories were combined (e.g., Black names with pleasant; White names with unpleasant); (d) a 40-trial block with the same content and category combination as the previous block; (e) a 20-trial block in which participants again categorized names only, with a reversal in the positions of Black–White; (f) a 20-trial block in which participants categorized all stimuli with the reversed positions of the Black–White categories (e.g., Black names paired with unpleasant; White names with pleasant); and, (g) a 40-trial block with the same content and category combinations as the previous block. The order of the combined blocks was counterbalanced.

Next came an explicit measure of attitudes toward Blacks, the Attitude Toward Blacks scale (Brigham, 1993). Items include, “Generally Blacks are not as smart as Whites” (reverse-coded) and “Black and White people are inherently equal.” Participants rated their agreement on a scale from 1 (*strongly disagree*) to 7 (*strongly agree*; $\alpha = .96$).

The next part provided both tests’ results. Participants in the self condition first saw a spinning wheel animation and a warning that the system was computing their scores. Participants in the others condition had no such delay. All participants read the following sentences (text that varied across conditions is shown inside square brackets): “The two tasks you have just completed were two measures of prejudice. This study is about your beliefs and impressions about these tasks. [Your score /Other people’s scores] on the first task (sorting task) [revealed/usually reveal] moderate to strong prejudice against Blacks. [Your score/Other people’s scores] on the second task (questionnaire) [revealed/usually reveal] slight to no prejudice against Blacks.”

After results were revealed, perceived validity was measured. Participants read, “We are interested in knowing your opinion about how accurately each of these tasks measures people’s real prejudice levels.” The perceived validity items were, “How well does the sorting task measure [your / other people’s] true prejudice levels?” and “How well does the questionnaire task measure [your/other people’s] true prejudice levels?” using a scale from 1 (*very poorly*) to 7 (*very well*). The first item measured the IAT’s

Table 1
Means (and Standard Deviations) of Perceived Validity and Individual Differences measured, per Experimental condition, Experiment 1

Variable	Condition	
	Self ($n = 24$)	Others ($n = 28$)
IAT perceived validity	1.92 (1.32)	3.46 (1.97)
IAT score	−0.48 (0.41)	−0.65 (0.26)
ATB perceived validity	5.58 (1.86)	5.11 (1.50)
ATB score	5.69 (1.17)	5.66 (1.43)

Note. IAT = Implicit Association Test; ATB = Attitude Towards Blacks Scale.

perceived validity and the second item the perceived validity of the explicit attitude measure. Participants were then debriefed and completed demographic questions.

Ethics statement. The present research was approved by the Ethics Committee of the host institution of the corresponding author.

Results

Table 1 displays descriptive information, Figure 1 shows a graphical representation of perceived validity means. Negative scores in the IAT represent a racial preference for Whites over Blacks.

Two multiple linear regressions were conducted—the first with perceived validity of the IAT as the dependent variable, the second with perceived validity of the Attitude Toward Blacks Scale as the dependent variable. In both models, the experimental manipulation (source of result) was effect-coded (self = −0.5, others = 0.5) and implicit attitude and explicit attitude were centered. To account for the influence of the implicit and explicit attitude on our experimental manipulation, we added interactions between implicit and explicit attitude and the experimental manipulation for a total of five parameters.

When perceived validity of the IAT was the dependent variable, the regression model was significant, $F(5, 46) = 5.31, p = .001$, adjusted $R^2 = .30$. The main effect of source of IAT was significant, $b = 1.48, SE = 0.45, 95\%$ confidence interval (CI) [0.58, 2.38], $p = .002$, such that participants in the others condition perceived the (negative) result to be more valid than participants in the self condition. The effect of explicit attitude (as measured by the Attitudes Toward Blacks Scale) was also significant, $b = −0.54, SE = 0.18, CI [−0.90, −0.19], p = .004$, such that, as explicit attitudes became more positive (i.e., increased), perceptions of validity of the IAT, which was said to reveal high prejudice, decreased. No other main effect or interaction was significant (all regression coefficients can be found in Table 2).

² The procedure used in Experiments 1–2 does not include two modern improvements in the application of the IAT: a larger number of the trials in the 5th block, where the keys used to categorize the targets of the IAT first change positions and which decreases the impact of block order (Nosek et al., 2005) and the use of more balanced stimuli, as in the original word set, White names were more familiar than Black names (Ottaway, Hayden, & Oakes, 2001).

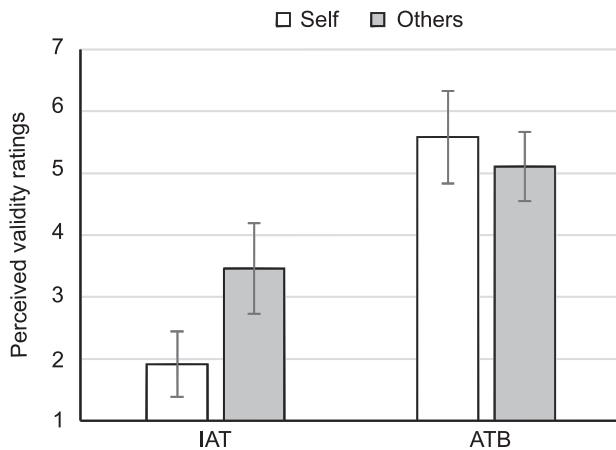


Figure 1. Mean perceived validity (with 95% confidence intervals) for the Implicit Association Test (IAT) and for the Attitude Toward Blacks scale (ATB), per condition (Experiment 1). To allow correct comparison of between-participants conditions (self vs. others), the CIs for the within-participants conditions (IAT and ATB) are not corrected.

Given the significant effect of source, and to directly test the self-other asymmetry hypothesis, we then analyzed the difference between the self and others conditions. As expected, participants in the self condition ($M = 1.92$, $SD = 1.32$) perceived the IAT as less valid than participants in the others condition ($M = 3.46$, $SD = 1.97$), $t(47.35) = -3.37$, $p = .002$, $d = -0.94$, 95% CI $[-1.51, -0.36]$.

When perceived validity of the Attitude Toward Blacks Scale was the dependent variable, the regression model was not significant, $F(5, 46) = 2.08$, $p = .086$, adjusted $R^2 = .10$. The regression coefficients for this model can be found in Table 2.

Discussion

People viewed the IAT as a more valid instrument for revealing implicit bias when it was said to reveal prejudice in others than when it was said to reveal prejudice in themselves.

People's explicit attitudes predicted perceptions of the IAT's validity, such that more positive explicit attitudes corresponded to lower perceptions of the IAT's validity (as the IAT results suggested high prejudice). Even so, when explicit (and implicit) attitudes were statistically taken into account, the predicted self-other difference in the IAT's perceived validity held.

Experiment 2

Whereas Experiment 1's IAT result always suggested high prejudice toward Blacks, Experiment 2 manipulated the desirability of the IAT result by suggesting either low or high bias (for the self or others). At this point, a question emerges as to whether the expected interaction pattern is symmetric, with equivalent self-other differences (in opposite directions) for a result suggesting high or low bias, or whether it is asymmetric such that there are larger self-other differences for results suggesting high degrees of bias.

Indeed, some studies have found self-other differences mostly when the domain in question is framed in a negative way (e.g., people believe that they are less likely than others to commit immoral actions), and not so much for positive domains (e.g., performing moral actions; Klein & Epley, 2016, 2017). Other studies, however, have found that people do consider themselves to be more virtuous than others (e.g., more cooperative, egalitarian, and norm-abiding; Alicke, 1985; Brown, 2012; Codol, 1975), and not simply less bad. Therefore, the following studies tested whether a self-other difference holds only for negative results, or whether it holds equally for positive and negative results, such that a low prejudice score would be seen as more valid for oneself than others and a high prejudice score would be seen as more valid for others than for oneself.

Method

Participants. Seventy-eight participants, recruited through MTurk, completed the experiment (32 women, 46 men, $M_{age} =$

Table 2
Regression Coefficients for Perceived Validity of the Implicit Association Test (IAT) and the Attitude Towards Blacks Scale (ATB), Experiment 1

Parameter	B	SE	95% CI	β	p
Perceived validity of the IAT					
Constant	2.62	.22	[2.17, 3.07]		<.001
IAT	-.38	.71	[-1.82, 1.05]	-.07	.593
ATB	-.54	.18	[-.90, -.19]	-.38	.004
Source of IAT result	1.48	.45	[.58, 2.38]	.40	.002
Source \times IAT	-1.71	1.43	[-4.59, 1.17]	-.16	.237
Source \times ATB	-.05	.36	[-.77, .67]	-.02	.891
Perceived validity of the ATB					
Constant	5.33	.23	[4.87, 5.80]		<.001
IAT	.45	.73	[-1.02, 1.93]	.09	.538
ATB	.05	.18	[-.32, .41]	.04	.797
Source of IAT result	-.40	.46	[-1.32, .52]	-.12	.389
Source \times IAT	-.05	1.46	[-3.00, 2.90]	-.01	.972
Source \times ATB	-1.04	.37	[-1.77, -.31]	-.40	.007

Note. CI = confidence interval.

Table 3
Means (and Standard Deviations) of Perceived Validity and Implicit Attitude, per Experimental Condition, Experiment 2

Variable	Condition			
	High prejudice		Low prejudice	
	Self (<i>n</i> = 16)	Others (<i>n</i> = 17)	Self (<i>n</i> = 20)	Others (<i>n</i> = 21)
IAT perceived validity	2.25 (2.02)	4.53 (1.33)	5.40 (1.54)	3.24 (1.34)
IAT score	-.58 (.31)	-.46 (.43)	-.52 (.33)	-.72 (.32)

Note. IAT = Implicit Association Test.

33.12, $SD_{age} = 10.26$, age range: 19–72 years).³ Sample size was not determined by a power analysis, but rather by a simple a priori rule of 20 participants per condition. Participants were paid \$1 for an estimated duration of 15 min (actual average 8:27 min). Fifteen additional participants started but did not complete the experiment, all dropping out before reaching the point where conditions differed. The requirements for participation were the same as in Experiment 1. Fifty-nine participants identified as White or Caucasian, six as Hispanic or Latino, four as Asian, four as mixed race, three as Black or African American, and two failed to report their ethnicity.

Four participants were excluded from analyses for having RTs under 300 ms in 10% or more of the critical trials of the IAT, as in Experiment 1. After exclusions, our statistical analysis had 80% power to detect effect sizes of $f^2 = .22$ (between medium = .15, large = .35; for a regression with 7 parameters and 74 participants), which is adequate to test the main hypothesis concerning the self-other asymmetry (as stated before, better-than-average effects are usually large; e.g., Hahn et al., 2014; Heine & Hamamura, 2007) but might be insufficient to test the impact of implicit attitude (more about this in the final meta-analysis and in the General Discussion).

Materials and procedure. Participants started by completing the same racial Black-versus-White IAT used in Experiment 1, which was described as a word categorization task. Then, participants in the self condition saw a spinning wheel animation accompanied by a warning to wait while the system computed their scores. Participants in the others condition experienced no such delay.

The IAT result information came next (text that varied between conditions is shown inside squared brackets): “The task you have just completed is a measure of prejudice. This study is about your beliefs and impressions about this task. [Your score/Other people’s scores] on this task [revealed/usually reveal] [low to no/moderate to strong] prejudice against Blacks. We are interested in knowing your opinion about how accurately this task measures [your/other people’s] real prejudice levels.”

Next came the measure of perceived test validity. The text read, depending on condition, “How well does the task measure [your/other people’s] true prejudice levels?” using a scale from 1 (*very poorly*) to 7 (*very well*). Last, participants were debriefed and provided demographic information.

Results

Table 3 displays descriptive information, Figure 2 shows a graphical representation of perceived validity means. Negative

scores in the IAT again represent a racial preference for Whites over Blacks.

A multiple linear regression was performed using perceived validity of the IAT as the dependent variable and source of IAT result (self = -0.5, others = 0.5), level of prejudice as suggested by the IAT (high prejudice = -0.5, low = 0.5), and implicit attitude as predictors. Implicit attitude, as measured by the IAT, was centered, and to control for its effect, we included all interactions between implicit attitude and conditions, for a total of 7 parameters.

The regression model was significant, $F(7, 66) = 6.58$, $p < .001$, adjusted $R^2 = .35$. The effect of level of prejudice was significant, $b = 1.03$, $SE = 0.38$, 95% CI [0.27, 1.78], $p = .009$, as was the interaction between level of prejudice and source of IAT result, $b = -4.14$, $SE = 0.76$, CI [-5.65, -2.62], $p < .001$. No other interaction or main effect was significant (all regression coefficients can be found in Table 4).

Given the significant interaction between source of IAT result and level of prejudice, and to directly test our self-other asymmetry hypothesis, we tested the difference between validity ratings for self and others at each level-of-prejudice condition. In the high prejudice condition, participants in the self condition ($M = 2.25$, $SD = 2.02$) perceived the IAT as less valid than did those in the others condition ($M = 4.53$, $SD = 1.33$), $t(31) = -3.86$, $p = .001$, $d = -1.34$, 95% CI [-2.09, -0.58]. In the low prejudice condition, participants in the self condition ($M = 5.40$, $SD = 1.54$) perceived the IAT as more valid than did those in the others condition ($M = 3.24$, $SD = 1.34$), $t(39) = 4.81$, $p < .001$, $d = 1.50$, CI [0.80, 2.19].

Discussion

Experiment 2 provided evidence for an asymmetry in judgments of the IAT’s validity. Validity perceptions varied depending on whether the IAT result pertained to people’s own bias or others’ bias, as well as whether it suggested a strong degree of bias. Replicating Experiment 1’s results, when the IAT suggested high prejudice, people judged it as less valid when it applied to them-

³ A total of 80 participants were recruited, yet two MTurkers submitted invalid completion codes. Because our informed consent did not explicitly mention that participants who did not reach the end of the experiment would not receive a reward and we could not confirm whether these MTurkers had started the experiment but dropped out because of our anonymity procedure, we opted for not denying payment nor replacing these participants.

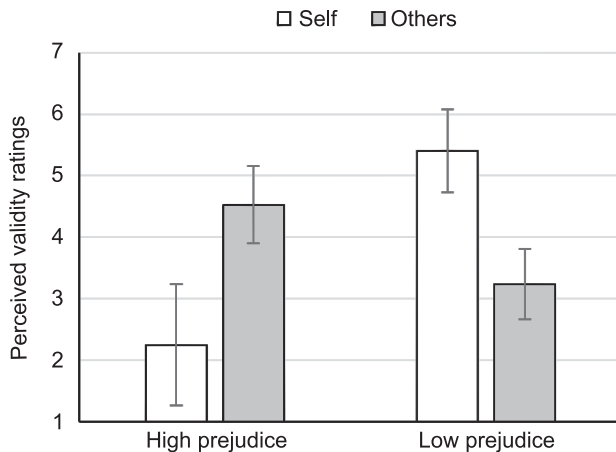


Figure 2. Mean Implicit Association Test perceived validity (with 95% confidence intervals), per condition (Experiment 2).

selves than when it referred to others. Furthermore, the opposite pattern was observed when the IAT suggested low prejudice. Then, people judged the IAT to be less valid when it applied to others than when it referred to themselves. In line with the results of Experiment 1, this self-other difference was not explained by people's implicit attitudes. These results go beyond Experiment 1 in showing that people do not invariably discredit the implications of their IAT results, but rather assess them depending on how desirable the result is.

Experiment 3

The third experiment aimed to make several advances. First, it sought to test whether the results of the previous experiments would generalize beyond Black/White ethnic groups to other social groups, namely elderly people.

Second, it sought to replicate the previous results while varying several methodological aspects of the previous experiments. A single-target IAT was used instead of the traditional two-target IAT, as it is not clear what the appropriate comparison group should be for the elderly category (adults, youth, or children). Moreover, we sampled Portuguese undergraduates who completed the experiment in the lab rather than American participants completing the experiment online. MTurk samples are of acceptable quality (e.g., Buhrmester, Kwang, & Gosling, 2011; Crump, McDonnell, & Gureckis, 2013; Paolacci, Chandler, & Ipeirotis, 2010), yet some issues have been raised, such as nonnaivety with popular research paradigms (Chandler, Mueller, & Paolacci, 2014).

The third goal involved assessing people's awareness of their implicit biases. Participants predicted their IAT result after taking the test and before receiving information about their own or others' results. Although Experiments 1–2 found no effect of participants' IAT scores on their perceptions of the IAT's validity, people's perception of their implicit bias may be a better predictor of the perceived validity of an IAT result than their IAT score.

As in Experiment 1, we measured participants' explicit attitudes to test whether the predicted self-other asymmetry would hold when controlling for them. As in Experiment 2, people received

IAT results suggesting either high or low bias referring either to themselves or others.

Method

Participants. Participants were 91 Portuguese undergraduates (79 women, 12 men, $M_{age} = 22.33$, $SD_{age} = 8.46$, age range: 17–53 years. Sample size was determined by the size of the sample available to the corresponding author in the semester in which the experiment ran in the laboratory. Using the same exclusion criteria as in previous experiments, four participants were excluded for having RTs under 300 ms in 10% or more of the critical trials of the IAT. After exclusions, our analysis had 80% power to detect effect sizes with $f^2 = .26$ (between medium = .15, large = .35; for a regression with 15 parameters and 87 participants). Again, this is probably adequate to test our self-other asymmetry hypothesis, but perhaps not sufficient to test the potential impact of individual differences, which may have smaller effects.

Materials and procedure. All materials were presented in Portuguese. Participants started by answering demographic questions. The order of the IAT and explicit measure was counterbalanced.⁴

The single-target IAT (Bluemke & Friese, 2008) involved three blocks of trials: (a) a 20-trial block in which participants categorized five pleasant and unpleasant pictures from the International Affective Picture System (Lang, Bradley, & Cuthbert, 2008); (b) a 35-trial block in which they categorized the pleasant and unpleasant pictures as well as five pictures of elderly people with neutral expressions taken from the FACES database (Ebner, Riediger, & Lindenberger, 2010), and where the label *elderly people* was paired with the *pleasant* label; and (c) a 35-trial block similar to the second block, but where the *elderly* label was paired with the *unpleasant* label. To bring the number of right- and left-hand responses closer (see Bluemke & Friese, 2008), the 35 trials of the combined blocks consisted of 10 trials with the pictures of elderly people, 10 trials of the valence label associated to the elderly people label, and 15 trials of the valence label not associated to the elderly people label. The sequence of target and attribute stimuli was fixed (target trial, attribute trial, target trial, attribute trial, attribute trial, repeat). This IAT also did not include trial feedback, and D_4 was used in calculating IAT scores.

In the explicit measure, participants used a 1 (*nothing*) to 7 (*a lot*) scale to rate five positive feelings (admiration, affection, appreciation, consideration, sympathy) and five negative feelings (hate, hostility, aversion, contempt, superiority) in the context of the sentence, "To what extent do you feel ____ towards elderly people?"

Regardless of order condition, participants completed a bias-awareness task (adapted from Hahn et al., 2014) immediately after the IAT. The text read: "The task that you have just finished is a version of the Implicit Association Test, which is used to measure attitudes, in this case regarding elderly people. This is done by comparing your answers in the block where the category elderly people was paired with the category pleasant with your answers in the block where the category elderly people was paired with the category unpleasant. We ask you to predict what your score will

⁴ There was no difference between the two orders of attitude instrument, $t(73.04) = 0.14$, $p = .887$, so we do not discuss order effects.

Table 4
Regression Coefficients for Perceived Validity of the Implicit Association Test (IAT),
Experiment 2

Parameter	<i>B</i>	<i>SE</i>	95% CI	β	<i>p</i>
Constant	3.87	.19	[3.49, 4.24]		<.001
Implicit attitude	1.01	.56	[-.10, 2.13]	.19	.073
Source of IAT result	.12	.38	[-.64, .88]	.03	.757
Level of prejudice	1.03	.38	[.27, 1.78]	.27	.009
Source \times Level	-4.14	.76	[-5.65, -2.62]	-.54	<.001
Source \times Implicit	-.17	1.11	[-2.40, 2.05]	-.02	.877
Level \times Implicit	-.10	1.11	[-2.33, 2.12]	-.01	.926
Source \times Level \times Implicit	1.05	2.23	[-3.40, 5.49]	.05	.640

Note. CI = confidence interval.

show.” Participants provided their estimates using the item: “I predict that the IAT result will reveal that my implicit attitude towards elderly people is . . .” using a 7-point scale ranging from 1 (*very negative*) to 7 (*very positive*).

Afterward, participants read the following sentences (text that varied between condition is shown inside square brackets): “In this part of the study, we are interested in your opinion about the Implicit Association Test, this is, the image categorization task you have done in an early part of this study. [Your score/Other people’s scores] on the Implicit Association Test [revealed that your/usually reveals that other people’s] level of prejudice against elderly people is [low/high]. We would like to know your opinion about how precise this task is in measuring [your/other people’s] real prejudice level.” Participants then rated their agreement with the item, “To what extent do you agree that the Implicit Association Test measures [your/other people’s] real prejudice level?” using a 7-point scale ranging from 1 (*strongly disagree*) to 7 (*strongly agree*). At end of the experiment, participants were debriefed.

Results

Table 5 displays descriptive information, Figure 3 shows a graphical representation of perceived validity means. Negative scores in the single-target IAT represent negative attitudes toward elderly people.

A multiple linear regression was performed using perceived validity of the IAT as the dependent variable, source of IAT result (self = -0.5, others = 0.5), prejudice level suggested by the IAT (high prejudice = -0.5, low = 0.5), implicit attitude, explicit attitude, and implicit attitude awareness as predictors. The individual-difference variables were centered, and we included all

interactions of each individual-difference variable with the experimental manipulations, resulting in a model with 15 parameters.

The regression model was significant, $F(15, 71) = 8.02$, $p < .001$, adjusted $R^2 = .55$. There was a significant effect of prejudice level, $b = 1.88$, $SE = 0.29$, 95% CI [1.29, 2.46], $p < .001$, and a significant interaction between source of IAT and prejudice level, $b = -3.32$, $SE = 0.59$, CI [-4.49, -2.15], $p < .001$. The only other significant effect was the main effect of implicit attitude awareness, $b = 0.36$, $SE = 0.12$, 95% CI [0.12, 0.60], $p = .004$, such that the more participants expected a positive attitude (i.e., low prejudice), the more they considered the IAT to be a valid measure of attitudes. All other coefficients were nonsignificant (see Table 6).

We then tested the difference between validity ratings for self and others at each prejudice level condition. Replicating the results of Experiment 2, in the high prejudice condition, participants perceived the IAT to be less valid for self ($M = 2.21$, $SD = 1.41$) than for others ($M = 3.92$, $SD = 1.20$), $t(48) = -4.64$, $p < .001$, $d = -1.31$, 95% CI [-1.92, -0.69], whereas in the low prejudice condition this pattern reversed (self: $M = 5.89$, $SD = 1.08$ vs. others: $M = 4.32$, $SD = 1.53$), $t(32.40) = 3.63$, $p = .001$, $d = 1.19$, CI [0.48, 1.89].

Discussion

Experiment 3 made several methodological changes to the prior experiments. Chiefly, Experiment 3 tested elderly people as the attitudinal social group, using European young adults as participants. Despite these changes, the results replicated Experiments 1–2’s findings. We once again observed a self-other asymmetry, in that people regarded the IAT as more valid for self than for others

Table 5
Means (and Standard Deviations) of Perceived Validity of the Implicit Association Test (IAT)
and Individual Differences Measured, per Experimental condition, Experiment 3

Variable	Condition			
	High prejudice		Low prejudice	
	Self ($n = 24$)	Others ($n = 26$)	Self ($n = 18$)	Others ($n = 19$)
IAT perceived validity	2.21 (1.41)	3.92 (1.20)	5.89 (1.08)	4.32 (1.53)
IAT score	-.16 (.22)	-.13 (.33)	.08 (.37)	-.17 (.41)
Explicit attitude	5.65 (.79)	5.84 (.79)	6.02 (.52)	5.94 (.49)
Implicit attitude awareness	5.21 (1.25)	4.81 (1.55)	5.56 (1.10)	5.37(1.21)

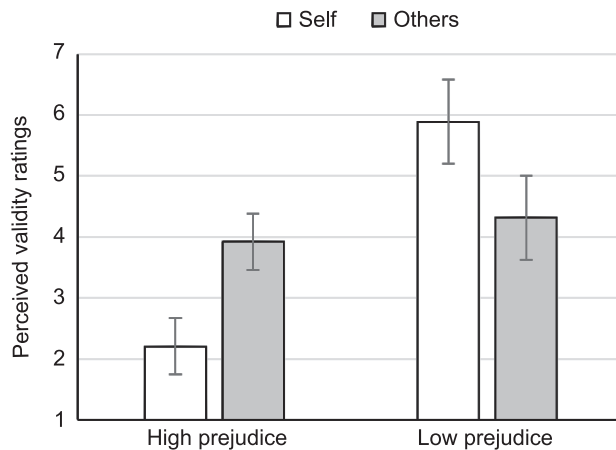


Figure 3. Mean Implicit Association Test perceived validity (with 95% confidence intervals), per condition (Experiment 3).

when its results indicated a positive attitude, whereas the opposite happened when test results were negative. These results were not affected by implicit attitudes, explicit attitudes, or awareness of implicit attitudes.

Experiment 4

Experiment 4 aimed to create a conservative test of our motivated reasoning account. It manipulated how undesirable it is to be implicitly biased by varying whether the social group in question is one for which social norms proscribe or prescribe negative attitudes: racists versus Blacks (Crandall et al., 2002). If the IAT's perceived validity is a function of how undesirable it is to seem biased, then participants should consider the IAT to be valid, and even more valid for themselves than for others, when it suggests that they are biased against a group that is generally despised in society, namely racists.

In addition, Experiment 4 measured participants' internal and external motivations to respond without prejudice (Plant & Devine, 1998). On the one hand, participants scoring high on the External Motivation Scale, who tend to admit feeling and thinking in a more prejudiced manner (Plant & Devine, 1998), may be more likely to accept a socially undesirable IAT result. On the other hand, participants scoring high on the Internal Motivation Scale report lower discrepancies between the way they believe they ought to think and feel and the way they think and feel in terms of racial bias (Plant & Devine, 1998), and may therefore feel threatened by a mismatch between their explicit attitude and their IAT result and consequently derogate the IAT by giving it lower validity ratings. We included these individual-difference variables to test whether the predicted self-other asymmetry would hold when accounting for them.

Mirroring Experiment 3, people reported their explicit attitudes and estimates of IAT results. We retested the hypothesis that people base their judgments of the IAT's validity, at least in part, on their own explicit attitudes and expectations of the result, with potential implications for the self-other asymmetry.

Method

Participants. Sample size was benchmarked off the smallest self-other asymmetry of previous experiments (d 's = 0.94–1.50). Accordingly, we aimed for 80% power to find an effect size of 0.90, which resulted in 21 per cell that was then rounded to 25 per cell. Thus, we recruited 200 participants (104 women, 96 men, $M_{\text{age}} = 38.72$, $SD_{\text{age}} = 12.18$, age range: 18–76 years) through MTurk, for which they were paid \$0.67 for an estimated duration of 10 min (real average 7 min). An extra 40 participants started the experiment but failed to complete it. In a loglinear analysis, no condition or interaction of conditions influenced attrition, all respective partial χ^2 with $ps > .150$. The requirements for participation were the same as in Experiments 1–2. Participants were not asked about their ethnicity.

Table 6
Regression Coefficients for Perceived Validity of the Implicit Association Test (IAT),
Experiment 3

Parameter	<i>B</i>	<i>SE</i>	95% CI	β	<i>p</i>
Constant	4.02	.15	[3.73, 4.31]		<.001
Implicit attitude	.05	.45	[−.84, .93]	.01	.918
Explicit attitude	−.28	.26	[−.80, .23]	−.11	.277
Implicit attitude awareness	.36	.12	[.12, .60]	.26	.004
Source of IAT result	.18	.29	[−.41, .76]	.05	.549
Level of prejudice	1.88	.29	[1.29, 2.46]	.51	<.001
Source × Level	−3.32	.59	[−4.49, −2.15]	−.46	<.001
Source × Implicit	−1.21	.89	[−2.99, .57]	−.11	.178
Source × Explicit	−.84	.52	[−1.87, .19]	−.16	.109
Source × Implicit Awareness	.44	.24	[−.04, .92]	.16	.072
Level × Implicit	−.45	.89	[−2.23, 1.32]	−.04	.613
Level × Explicit	.56	.52	[−.48, 1.59]	.10	.287
Level × Implicit awareness	.30	.24	[−.18, .79]	.11	.213
Source × Level × Implicit	1.80	1.78	[−1.75, 5.36]	.08	.315
Source × Level × Explicit	−1.79	1.03	[−3.85, .27]	−.17	.088
Source × Level × Implicit Awareness	.66	.48	[−.30, 1.63]	.12	.175

Note. CI = confidence interval.

Using the same exclusion criteria as in previous experiments, 10 participants were removed (RTs <300 ms in $\geq 10\%$ of critical IAT trials). After exclusions, the main analysis had 80% power to detect effect sizes with $f^2 = .15$ (which is the cutoff for medium effect-sizes; for a regression with 31 parameters and 190 participants). Once again, this is probably sufficient to test the main hypothesis regarding the self-other asymmetry but might still be too low to test the impact of the individual-difference variables.

Materials and procedure. Participants started by completing a single-target IAT similar to the one used in Experiment 3, in one of two versions. We selected two groups, racists and Blacks, as they vary in how acceptable it is to discriminate against them (Crandall et al., 2002). To represent Blacks, we used four public domain images of Blacks, cropped as described in Nosek et al. (2007). To represent racists, we used two pictures of young men in front of Nazi flags and two pictures of Ku Klux Klan members in uniform, taken from the International Affective Picture System, and from pictures with public domain licenses. For the pleasant and unpleasant pictures, we again used pictures from the International Affective Picture System.

After the IAT, participants performed the bias-awareness task (adapted from Hahn et al., 2014), similar to Experiment 3. Thus, they first read the following (the text within square brackets varied depending on condition): “The sorting task you have just completed is a version of the Implicit Association Test (IAT) and is used to measure attitudes, in this case, towards [Blacks/racists]. This is done by comparing your answers in the block where the category [Blacks/racists] was paired with the category pleasant with the block where the category [Blacks/racists] was paired with the category unpleasant”. Participants gave bias awareness ratings using the items, “I predict that the IAT comparing my reactions when the category [Blacks/racists] was in the same key as pleasant vs. when [Blacks/racists] was in the same key as unpleasant will show that my attitude is . . .” using a scale from 1 (*strongly negative toward* [Blacks/racists]) to 7 (*strongly positive toward* [Blacks/racists]).

Participants then saw the pictures of the target group again (Blacks or racists, depending on condition) and rated their explicit attitude toward them, “How favorable or unfavorable is your attitude towards the person(s) in this picture?” using a Visual Analogue Scale ranging from *very unfavorable* to *very favorable* (no numerical scale anchors were displayed; the system recorded answers on a 1–100 scale).⁵

Afterward, participants read the following sentences: “The picture categorization task (Implicit Association Test) you did in the beginning of this study is a measure of attitudes. This study is about your beliefs and impressions about that task. [Your score/Other people’s scores] in this task [revealed/usually reveal] a [positive/negative] attitude towards [Blacks/racists]. We are interested in knowing your opinion about how accurately this task measures [your/other people’s] real attitudes towards [Blacks/racists].” Then participants responded to the question, “How well does the task measure [your/other people’s] true attitudes towards [Blacks/racists]?” using a scale ranging from 1 (*very poorly*) to 7 (*very well*).

In the Black target group condition, participants completed the External and the Internal Motivation to respond without prejudice Scales (Plant & Devine, 1998).⁶ Last, participants were debriefed and answered demographic questions.

Results

Table 7 displays descriptive information, Figure 4 shows a graphical representation of perceived validity means. Negative scores in the IAT represent a negative attitude toward Blacks or racists, depending on condition.

A multiple regression was performed with perceived validity of the IAT as the dependent variable, and source of IAT result (self = -0.5 , others = 0.5), valence of attitude suggested by the IAT (positive = 0.5 , negative = -0.5), attitudinal target of the IAT (racists = -0.5 , Blacks = 0.5), implicit attitude, explicit attitude, and implicit attitude awareness as predictors. We centered individual differences and added interactions between each individual difference and our experimental manipulations, resulting in a total of 31 parameters. The scores on the Internal and External Motivation Scales were not included, as they did not result in significant effects or interactions in a regression with perceived validity of the IAT in the Blacks condition (the condition where these scales were applied⁶).

The regression model was significant, $F(31, 158) = 9.64$, $p < .001$, adjusted $R^2 = .59$. The interaction between valence of attitude and source of IAT result was significant, $b = -2.95$, $SE = 1.45$, 95% CI [$-5.81, -0.09$], $p = .043$, whereas the three-way interaction between the three experimental manipulations was not significant, $b = -1.23$, $SE = 2.90$, CI [$-6.95, 4.49$], $p = .672$. There was a significant interaction between the three experimental manipulations and explicit attitude, $b = 0.20$, $SE = 0.10$, CI [$0.01, 0.39$], $p = .038$. The only remaining significant effect was the interaction between the valence of attitude manipulation and explicit attitude, $b = 0.05$, $SE = 0.02$, CI [$0.00, 0.10$], $p = .049$. All coefficients can be found in Table 8.

Yet, severe multicollinearity was found in the regression model, as the attitudinal target manipulation correlated strongly with participants’ explicit attitudes, $r(188) = .88$, $p < .001$, implicit attitude, $r(188) = .51$, $p < .001$, and implicit attitude awareness, $r(188) = .65$, $p < .001$. Thus, clear conclusions on the role of the covariates cannot be borne out in this experiment. A regression model without the covariates (see all coefficients in Table 9) was significant, $F(7, 182) = 26.33$, adjusted $R^2 = .48$, and supported the Source \times Valence \times Target three-way interaction, $b = -5.92$, $SE = 0.91$, 95% CI [$-7.72, -4.12$], $p < .001$.

We then tested the self-other asymmetry for each combination of valence and target. When the IAT revealed a negative attitude toward Blacks, participants in the self condition ($M = 2.36$, $SD = 1.66$) perceived the IAT to be less valid than did participants in the others condition ($M = 3.52$, $SD = 1.58$), $t(48) = -2.53$, $p = .015$, $d = -0.72$, 95% CI [$-1.28, -0.14$]. When the IAT revealed a positive attitude toward Blacks, participants in the self condition ($M = 5.86$, $SD = 1.61$) perceived the IAT to be more valid than did participants in the others condition ($M = 3.65$, $SD = 1.75$), $t(43) = 4.41$, $p < .001$, $d = 1.31$, CI [$0.66, 1.95$]. When the IAT

⁵ All participants in this experiment also answered the previous question from the point of view of most other people (order was counterbalanced), a variable that was added to test an exploratory hypothesis relating to implicit attitudes and unrelated with perceived validity and therefore is not reported in the analysis.

⁶ Several items in these scales make little sense when applied to racists (e.g., “I try to act nonprejudiced towards [racists] because of pressure from others”). Therefore, we used this scale only in the condition where Blacks was the target group.

Table 7

Means (and Standard Deviations) of Perceived Validity of the Implicit Association Test (IAT) and Individual Differences Measured, per Experimental Condition, Experiment 4

Variable	Condition							
	Blacks				Racists			
	Positive attitude		Negative attitude		Positive attitude		Negative attitude	
	Self (<i>n</i> = 22)	Others (<i>n</i> = 23)	Self (<i>n</i> = 25)	Others (<i>n</i> = 25)	Self (<i>n</i> = 27)	Others (<i>n</i> = 23)	Self (<i>n</i> = 22)	Others (<i>n</i> = 23)
IAT perceived validity	5.86 (1.61)	3.65 (1.75)	2.36 (1.66)	3.52 (1.58)	1.81 (1.71)	2.57 (1.20)	6.41 (.91)	4.61 (1.85)
IAT score	-.20 (.41)	-.02 (.45)	-.03 (.49)	-.05 (.47)	-.54 (.31)	-.56 (.28)	-.49 (.32)	-.53 (.30)
Explicit attitude	65.48 (17.98)	65.37 (19.68)	64.06 (21.34)	65.05 (22.71)	8.78 (15.79)	4.09 (7.25)	4.92 (6.42)	4.89 (5.91)
Implicit attitude awareness	4.50 (1.26)	4.87 (1.42)	4.68 (1.31)	4.28 (1.49)	2.37 (1.64)	2.22 (1.17)	2.27 (1.39)	2.04 (1.30)
External motivation	4.03 (2.45)	3.86 (2.12)	4.14 (2.56)	4.05 (2.52)				
Internal motivation	7.69 (1.70)	7.60 (2.15)	6.53 (2.19)	6.64 (2.49)				

revealed a negative attitude toward racists, participants in the self condition ($M = 6.41$, $SD = 0.91$) perceived the IAT as more valid than did participants in the others condition ($M = 4.61$, $SD = 1.85$), $t(32.31) = 4.17$, $p < .001$, $d = 1.24$, $CI [0.60, 1.88]$. Finally, when the IAT revealed a positive attitude toward racists, participants in the self condition ($M = 1.81$, $SD = 1.71$) perceived the IAT to be less valid than participants in the others condition ($M = 2.57$, $SD = 1.20$), though this difference is only marginally significant, $t(48) = -1.77$, $p = .084$, $d = -0.50$, $CI [-1.06, 0.07]$.

Discussion

The key advance in Experiment 4 was to test desirability independently of valence of attitude revealed by the IAT result. Replicating prior experiments, when the social norm proscribed the expression of bias (i.e., when the target was Blacks), a negative attitude result was seen as less valid for the self than for others, whereas when the IAT revealed a positive attitude, the opposite was true. By contrast, when the social norm proscribed the expression of bias, this pattern reversed: When the IAT suggested a negative attitude toward racists, the IAT was considered more valid for self than others, whereas when the IAT pointed to a positive attitude toward racists it was considered less valid for self than others.

In the prior experiments, the desirability of the IAT result (e.g., that it is desirable to have a positive attitude toward Blacks) was confounded with the valence of the attitude the IAT revealed (i.e., that the attitude toward Blacks is positive). Experiment 4 provided a strong test of the desirability prediction by separating those factors. In the key condition that pits these factors against each other, people learned that their IAT score indicates that they harbor negative attitudes toward racists. If people's reactions in the prior studies are simply a function of being told they possess negative attitude toward minority groups, then this group should likewise perceive the test to lack validity. That is not what happened. Instead, people viewed the test as having validity, thereby supporting the hypothesis that the perceived validity of the IAT depends on the desirability of its results.

Experiment 5

Experiment 5's aim was to investigate whether the patterns of perceived validity observed in the previous experiments influence reactions to the use of the IAT as a personnel selection tool for jobs involving interacting with discriminated minorities. We reasoned that perceiving the IAT as a valid personnel selection tool (with the aim to select the least biased people) requires regarding it as a valid measure of implicit bias, much like perceiving an IQ test as

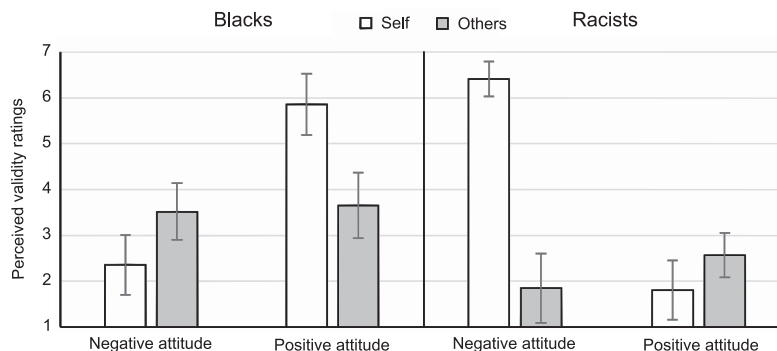


Figure 4. Mean Implicit Association Test perceived validity (with 95% confidence intervals), per condition (Experiment 4).

Table 8
Regression Coefficients for Perceived Validity of the Implicit Association Test (IAT), Model With Individual Differences, Experiment 4

Parameter	<i>B</i>	<i>SE</i>	95% CI	β	<i>p</i>
Constant	4.21	.36	[3.50, 4.92]		<.001
Source of IAT result	.34	.72	[-1.09, 1.77]	.08	.638
Valence of attitude	-1.15	.72	[-2.58, .28]	-.26	.115
Attitudinal target	-.59	.72	[-2.02, .84]	-.14	.417
Implicit attitude	-.50	.33	[-1.14, .14]	-.10	.125
Explicit attitude	.01	.01	[-.01, .04]	.20	.295
Implicit attitude awareness	.03	.10	[-.16, .22]	.03	.722
Source \times Valence	-2.95	1.45	[-5.81, -.09]	-.34	.043
Source \times Target	-2.39	1.45	[-5.25, .47]	-.28	.100
Source \times Implicit	-.02	.65	[-1.30, 1.26]	.00	.975
Source \times Explicit	.04	.02	[-0.01, .09]	.31	.096
Source \times Awareness	-.07	.19	[-.45, .31]	-.03	.702
Valence \times Target	.95	1.45	[-1.91, 3.81]	.11	.511
Valence \times Implicit	.66	.65	[-.62, 1.94]	.07	.310
Valence \times Explicit	.05	.02	[.00, .10]	.37	.049
Valence \times Awareness	.41	.19	[.03, .79]	.17	.033
Target \times Implicit	-.36	.65	[-1.65, .92]	-.03	.578
Target \times Explicit	-.03	.02	[-.08, .02]	-.11	.221
Target \times Awareness	.15	.19	[-.23, .53]	.05	.423
Source \times Valence \times Target	-1.23	2.90	[-6.95, 4.49]	-.07	.672
Source \times Valence \times Implicit	.32	1.30	[-2.25, 2.88]	.02	.809
Source \times Valence \times Explicit	-.08	.05	[-.18, .02]	-.31	.099
Source \times Valence \times Awareness	-.14	.38	[-.89, .62]	-.03	.726
Source \times Target \times Implicit	.64	1.30	[-1.93, 3.20]	.03	.625
Source \times Target \times Explicit	-.08	.05	[-.18, .01]	-.31	.094
Source \times Target \times Awareness	.58	.38	[-.18, 1.34]	.12	.135
Valence \times Target \times Implicit	-1.07	1.30	[-3.63, 1.50]	-.06	.412
Valence \times Target \times Explicit	.02	.05	[-.07, .12]	.09	.615
Valence \times Target \times Awareness	.13	.38	[-.63, .89]	.03	.728
Source \times Valence \times Target \times Implicit	-1.55	2.60	[-6.68, 3.58]	-.04	.551
Source \times Valence \times Target \times Explicit	.20	.10	[.01, .39]	.39	.038
Source \times Valence \times Target \times Awareness	-.29	.77	[-1.81, 1.23]	-.03	.707

Note. CI = confidence interval.

a valid personnel selection tool (with the aim to select the most intelligent people) depends on regarding it as a valid measure of intelligence. As such, we tested whether validity perceptions of the IAT mediated the effect of Desirability \times Source interaction on endorsement of the IAT as a selection tool in the workplace.

Method

Participants. One hundred eighteen participants, recruited through MTurk, completed the experiment (58 women, 60 men, $M_{\text{age}} = 36.79$, $SD_{\text{age}} = 12.68$, age range: 18–75 years).⁷ Sample size was not determined by a power analysis, but rather by a simple a priori rule of 30 participants per condition. Participants were paid \$0.45 for an estimated duration of 6 min (real average 5:04 min). An extra 25 participants started the experiment but dropped out, 23 of whom did so during the IAT. The two participants who dropped out after completing the IAT were both in the positive attitude feedback condition, one before feedback and the other during debriefing. Given that only 1.4% dropped out of the experiment after condition assignment, minimal selective attrition effects are expected. The requirements for participation were the same as in Experiments 1, 2, and 4. Using the same exclusion criteria as in all previous experiments, seven participants were removed for having RTs under 300 ms in 10% or more of the critical trials of the IAT.

Materials and procedure. The experiment began with the Blacks single-target IAT used in Experiment 4, which was introduced as a categorization task. Afterward, participants read the following sentences (text varying between conditions is shown inside squared brackets): “The picture categorization task (Implicit Association Test) you did in the beginning of this study is a measure of attitudes. This study is about your beliefs and impressions about that task. [Your score/Other people’s scores] in this task [revealed/usually reveal] a [positive / negative] attitude towards Blacks. We are interested in knowing your opinion about how accurately this task measures [your/other people’s] real attitudes towards Blacks.” Just below the sentences, participants were asked “How well does the task measure [your/other people’s] true attitudes towards Blacks?”, with participants answering using the same 7-point perceived validity scale as in Experiment 4 ranging from 1 (*very poorly*) to 7 (*very well*).

Then, participants read the following paragraph:

These days, many companies and organizations are spending a significant portion of their human resources budgets on measures both to increase the diversity of their workforce and to ensure that their

⁷ A total of 120 participants were recruited but, as in Experiment 2, two MTurkers submitted invalid completion codes. See Footnote 3.

Table 9
Regression Coefficients for Perceived Validity of the Implicit Association Test (IAT), Model Without Individual Differences, Experiment 4

Parameter	<i>B</i>	<i>SE</i>	95% CI	β	<i>p</i>
Constant	3.85	.11	[3.62, 4.07]		<.001
Source of IAT result	-.53	.23	[-.98, -.08]	-.12	.022
Valence of attitude	-.75	.23	[-1.20, -.30]	-.17	.001
Attitudinal target	.00	.23	[-.45, .45]	.00	.998
Source \times Valence	-.41	.46	[-1.31, .49]	-.05	.370
Source \times Target	.00	.46	[-.90, .90]	.00	.999
Valence \times Target	5.14	.46	[4.24, 6.04]	.59	<.001
Source \times Valence \times Target	-5.92	.91	[-7.72, -4.12]	-.34	<.001

Note. CI = confidence interval.

employees respect diversity within their institution and the diversity of their customers. The Implicit Association Test has been suggested, by some, to be a good tool to recruit and select individuals who will be unprejudiced and therefore more eligible to work with diverse colleagues and customers.

Below the paragraph, participants completed an adapted version of Chan, Schmitt, Sacco, and DeShon's (1998) Test Reaction Scale. All nine items of the scale were adapted by specifying that the skill or job involved interacting with Blacks and were answered in a 7-point scale from 1 (*strongly disagree*) to 7 (*strongly agree*). The Test Reaction Scale has three subscales that measure reactions to tests: face validity perceptions (e.g., "I can see a clear connection between the Implicit Association Test and what I think is required by a job that involves interacting with Blacks"), predictive validity perceptions (e.g., "I am confident that the test can predict how well an applicant will perform on a job that involves interacting with Blacks"), and fairness perceptions (e.g., "I feel that using the test to select applicants for a job that involves interacting with Blacks is fair"). The Test Reaction Scale revealed high internal consistency, $\alpha = .93$. Participants were then debriefed and completed demographic questions. They were not asked to report their ethnicity.

Results

Table 10 displays descriptive information, Figure 5 shows a graphical representation of perceived validity means. Negative IAT scores represent negative attitudes toward Blacks.

A multiple linear regression was conducted with perceived validity of the IAT as the dependent variable, and with source of IAT result (self = -0.5, others = 0.5), valence of attitude (positive = 0.5, negative = -0.5), and implicit attitude as predictors.

Implicit attitude was centered, and all interactions between the main effects were included, for a total of seven parameters. This regression model with seven parameters and a sample size of 111 participants had 80% power to detect an effect size of at least $f^2 = .14$ (slightly below the medium cutoff, .15).

The regression model was significant, $F(7, 103) = 10.15, p < .001$, adjusted $R^2 = .37$. The effect of valence of attitude was significant, $b = 2.12, SE = 0.30, 95\% CI [1.52, 2.72], p < .001$, as was the effect of source of IAT, $b = -0.67, SE = 0.30, CI [-1.27, -0.07], p = .029$, and the interaction between these two experimental manipulations, $b = -2.52, SE = 0.60, CI [-3.72, -1.32], p < .001$. No other effect or interaction was significant (all regression coefficients can be found in Table 11).

As in previous experiments, we then tested self-other differences for each valence condition. When the IAT revealed a positive attitude, participants in the self condition ($M = 5.46, SD = 1.66$) perceived the IAT as more valid than did participants in the others condition ($M = 3.53, SD = 1.76$), $t(56) = 4.27, p < .001, d = 1.13, CI [0.56, 1.68]$. When the IAT was said to reveal a negative attitude, the self-other asymmetry found in the prior studies was not significant, $t(51) = -1.52, p = .135, d = -0.42, 95\% CI [-0.96, 0.13]$.

We then tested whether the self-other asymmetry in the perceived validity of the IAT as a measure of attitudes spilled over to participants' responses to the idea of using the IAT as a tool for personnel selection. As a reminder, perceived validity of the IAT as a measure of attitudes was measured by the agreement to the sentence "How well does the task measure [your/other people's] true attitudes towards Blacks?", and perceived validity of the IAT as a personnel selection tool was measured using an adaptation of the Test Reaction Scale, which consists of questions relating to the

Table 10
Means (and Standard Deviations) of Perceived Validity of the Implicit Association Test (IAT), Test Reaction Scale Scores, and Implicit Attitude, per Experimental condition, Experiment 5

Variable	Condition			
	Negative attitude		Positive attitude	
	Self (<i>n</i> = 26)	Others (<i>n</i> = 27)	Self (<i>n</i> = 26)	Others (<i>n</i> = 32)
IAT perceived validity	2.08 (1.44)	2.67 (1.39)	5.46 (1.66)	3.53 (1.76)
Test Reaction Scale	2.31 (1.13)	2.66 (1.27)	3.06 (1.43)	2.75 (1.34)
IAT score	-.05 (.44)	-.06 (.35)	-.04 (.55)	-.07 (.46)

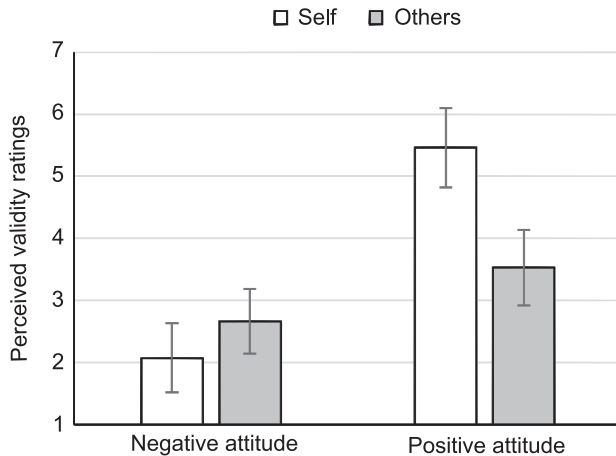


Figure 5. Mean Implicit Association Test perceived validity (with 95% confidence intervals), per condition (Experiment 5).

adequacy of a test (in this case, the IAT) in a given job (in this case, an adaptation was done so that the job was one that involved interacting with Blacks; e.g., “I am confident that the test can predict how well an applicant will perform on a job that involves interacting with Blacks”). For both types of measures, higher ratings indicated higher perceived validity.

Thus, we tested whether the interaction between desirability and source of IAT result had an indirect impact on reactions to using the IAT as a tool for personnel selection, mediated by the perceived validity of the IAT as an attitudinal measure. This amounts to testing a moderated mediation model, which we did using Model 8 of the macro PROCESS Procedure for SPSS (Hayes, 2013), with 10,000 bootstrap samples; see Figure 6 for the model definition and other details. The moderated mediation test was significant ($b = -1.02, SE = 0.31, 95\% CI [-1.73, -0.51]$). Thus, the Desirability \times Source interaction effect on perceived validity of the IAT as an attitudinal measure had a spill-over effect on perceived validity of the IAT as a personnel selection tool.

Given that perceived validity of the IAT as a personnel selection test was highly correlated with perceived validity of the IAT as a measurement of attitudes, $r(109) = .49, p < .001$, and that both were measured in short succession, we also considered an alternative moderated mediation model with perceived validity of the IAT as a measure of attitudes as the outcome and perceived validity of the IAT as a personnel selection tool as the mediator.

This alternative model was not supported ($b = -0.40, SE = 0.30, 95\% CI [-1.07, 0.15]$).

Discussion

In the fifth experiment, we replicated our findings from Experiments 2–4 regarding the effect of the interaction between source of IAT result and desirability of IAT result on the perceived validity of the IAT, although the self-other difference was only significant in the positive condition. More importantly, this effect carried over to people’s evaluations of the IAT as a personnel selection tool. A moderated mediation demonstrated that the influence of the desirability and source of IAT result on perceived validity of the IAT as a measure of attitudes had an indirect impact on the validity of the IAT as a personnel selection tool.

Experiment 6

The last experiment sought to make a stronger case for the self-enhancement (i.e., motivated reasoning) hypothesis by testing it against a competing self-verification account (Swann, 1983). A potential alternative explanation of the results thus far is that participants simply expect themselves to be egalitarian and others to be less so, and therefore IAT scores may be seen as more or less credible to the extent that they fit this expectation.

To test this alternative explanation, people were given IAT results that deviated from what they believed to be their own, or others’, explicit attitudes. In the undesirable condition, for example, if a participant indicated that he expected his attitude toward the target group to be “moderately positive” (the seventh point on a 9-point scale), the IAT result the participant would receive would be two points lower (i.e., the fifth point of the scale, meaning “neither positive nor negative”, thus indicating more bias than expected). In the desirable condition, if the participant reported that same attitude (i.e., the seventh point of the scale, meaning “moderately positive”), the IAT result would be two points higher (i.e., “extremely positive”, the ninth point of the scale, thus indicating less bias). If perceived validity is driven by congruency with expectations, then participants should find an IAT score to be equally invalid regardless of whether it is more positive or more negative than expected. If, however, desirability is the major determinant of perceived validity, then participants should find a score that is more positive than expected to be more valid than a score that is more negative than expected.

Experiment 6 also included several improvements: the IAT was again a standard IAT instead of a single-target IAT, which has

Table 11
Regression Coefficients for Perceived Validity of the Implicit Association Test (IAT), Experiment 5

Parameter	B	SE	95% CI	β	p
Constant	3.43	.15	[3.13, 3.73]		<.001
Implicit attitude	-.27	.36	[-.97, .44]	-.06	.457
Source of IAT	-.67	.30	[-1.27, -.07]	-.17	.029
Valence of attitude	2.12	.30	[1.52, 2.72]	.53	<.001
Source \times Valence	-2.52	.60	[-3.72, -1.32]	-.32	<.001
Source \times Implicit	-1.08	.71	[-2.50, .33]	-.12	.132
Valence \times Implicit	.62	.71	[-.80, 2.03]	.07	.390
Source \times Valence \times Implicit	.37	1.43	[-2.45, 3.20]	.02	.793

Note. CI = confidence interval.

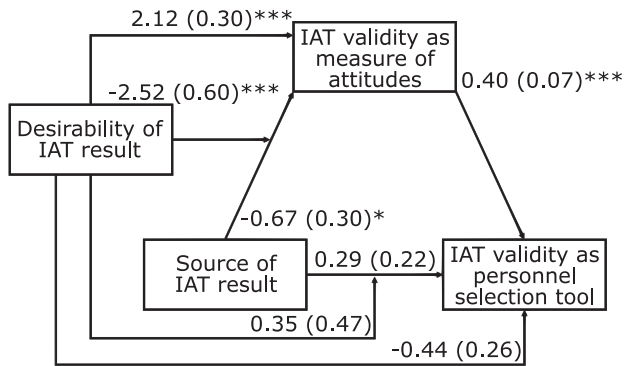


Figure 6. Diagram of the moderated mediation of Experiment 5. * $p < .05$. *** $p < .001$.

lower psychometric qualities (Bar-Anan & Nosek, 2014), and this time without the issues of Experiments 1–2 (see Footnote 2). Our measure of perceived validity was expanded to a four-item scale, including two reverse-coded items, to increase psychometric quality. In addition, sample size was increased, and an attention check was included.

Method

Participants. A power analysis was conducted in this experiment. Using G*Power 3.1, we estimated the necessary sample size for 80% power with a regression with 15 parameters (a saturated regression model) and an effect size between small and moderate (f^2 between 0.02 and 0.15, so 0.085), resulting in a total of 235 participants, which we rounded up to 240.

We aimed to recruit 240, and 242 participants successfully completed the experiment (119 women, 123 men, $M_{\text{age}} = 38.02$, $SD_{\text{age}} = 12.66$, age range: 20–79 years, 175 White or Caucasian, 24 as Black or African American, 17 Hispanic or Latino, 16 Asian, five mixed race, one Jewish, and one Native American). Participants were recruited through MTurk and were paid \$0.75 for an estimated duration of 10–15 min (real average 7:55 min). An extra 41 participants started but did not finish the experiment, 38 of whom dropped out before completing the IAT. As only three participants dropped out after condition assignment, selective attrition was not expected to impact the results. Requirements for participation were the same as for all our previous MTurk experiments.

Twenty-three participants were removed for having RTs under 300 ms in 10% or more of the critical trials of the IAT, as for all our previous experiments, and 35 additional participants failed the attention check (described below in the materials and procedure section) and therefore were removed from analysis. In total 184 participants remained. After exclusions, our statistical analysis had 80% power to detect effect sizes with $f^2 = .10$ (small effects fall between $f^2 = .02$ and $.15$; for a regression with 11 parameters and 184 participants).

Materials and procedure. Participants started by answering the demographic questions. The first part of the experiment consisted of collecting participants' implicit attitude as well as their explicit attitude (self condition) or their beliefs about other people's attitudes (others condition). Participants in the self condition

read: "Now, we ask you to express what you feel about Black people. Please complete the following sentence: My attitude towards Blacks is . . .", whereas participants in the others condition read, "Now, we ask you to infer what other people in general feel about Black people. Please complete the following sentence: Other people's attitudes towards Blacks are . . .". All participants answered using a fully labeled 9-point Likert scale ranging from 1 (*extremely negative*) to 9 (*extremely positive*). Only the verbal labels were presented to participants.

For the IAT, we used a White-versus-Blacks IAT created using iatgen (Carpenter et al., 2018). For stimuli, we used both the face picture set and valenced words list used by Nosek et al. (2007). Participants were given trial feedback and forced to correct their answers, so D_1 was computed using iatgen's Shiny app. A total of seven blocks, as in Experiment 1, were used, but this time with 40 trials in Block 5, as recommended by Nosek, Greenwald, and Banaji (2005). Order of IAT and self or others' explicit attitudes was counterbalanced.

As a transition to the second part of the experiment, participants read,

You have now finished the first part of this study. In reality, this study is about the categorization task in which you sorted pictures of faces into the Black and White categories and words into the positive and negative categories. This task is actually a measure of attitudes and this study is about your beliefs and impressions about such a task as a measure of attitudes.

In the second part of the experiment, participants in the self condition first saw a spinning wheel animation and were asked to wait a few seconds as the system computed their scores. All participants read the following sentences: "[Your result/Other people's results] in the categorization task [revealed/typically reveals] that [your/their] attitude towards Blacks is [X]. You stated that [your attitude/others' attitudes] towards Blacks [is/were] [Y]. Therefore, [your result/others' results] in this task [has revealed/typically reveal] [Z] the one you had stated." In which X was the verbal label of a deviation of 2 from the participants' own reported attitude (in the self condition) or expectation of others' attitudes (in the others condition). So, for example, if a participant stated that others' attitudes were "moderately negative" (3) and was assigned to the undesirable condition, the participant would see "extremely negative" (1) as others' typical IAT results, whereas if the participant was assigned to the desirable condition, the result would be "neither positive nor negative" (5). In cases where adding or subtracting two points to the Likert scale would lead the result to be outside of the scale's bounds, the extreme of the scale was displayed instead. So, if a participant chose "extremely positive" (9) and ended up in the desirable deviation condition, the participant would see "extremely positive" (9) as the IAT result. Y was a restatement of the participants' original answer to the 9-point Likert scale and Z was a verbal description of the difference between the participants' explicit attitude (of self or others) and the IAT result: either "a more negative attitude than," "a more positive attitude than" or "an equal attitude to."

Immediately below the text revealing the IAT result, the perceived validity scale was presented. This scale consisted of four items, the first of which was similar to the one we used in the previous experiments ("In your opinion, is this task a good measure of [your / other people's] true attitudes towards Blacks?")

Table 12
Means (and Standard Deviations) of Perceived Validity of the Implicit Association Test (IAT), Individual Differences Measured, and Observed Deviations of IAT Result from Expectation, per Experimental Condition, Experiment 6

Variable	Condition			
	Desirable deviation		Undesirable deviation	
	Self (<i>n</i> = 48)	Others (<i>n</i> = 43)	Self (<i>n</i> = 43)	Others (<i>n</i> = 50)
IAT perceived validity	5.18 (1.59)	4.66 (1.22)	3.32 (2.06)	4.42 (1.48)
IAT score	-.35 (.43)	-.42 (.41)	-.34 (.46)	-.33 (.46)
Explicit self/others	6.52 (1.91)	4.65 (1.79)	6.63 (1.89)	4.84 (1.62)
Deviation = 2	30	40	42	47
Deviation = 1	14	2	0	3
Deviation = 0	4	1	1	0

Note. The three deviation rows refer to the number of participants who saw 2, 1 or no difference between their stated self or others' attitude and the IAT result.

using a scale ranging from 1 (*extremely bad*) to 9 (*extremely good*) and three items adapted from Swann, Griffin, Predmore, and Gaines (1987): "How accurate do you think this task was in measuring [your / other people's] attitudes towards Blacks?", using a scale ranging from 1 (*extremely accurate*) to 9 (*extremely inaccurate*; reverse coded); "How much do you think your performance on this task could reveal about [your / other people's] attitudes towards Blacks?," using a scale ranging from 1 (*nothing at all*) to 9 (*a great deal*); "To what extent do you think the result of this task was a result of [your/other people's] attitudes towards Blacks?," using a scale ranging from 1 (*totally a result of [my/other people's] attitudes*) to 9 (*not at all a result of [my/other people's] attitudes*) (reverse coded).

The attention check was presented after these four items with the following text: "To what extent do you think that the result of the categorization task was not a result of [your/other people's] attitudes towards Blacks, but a result of factors not related to attitudes towards Blacks? This is an attention check, please give seven as an answer to this question," using a scale ranging from 1 (*totally a result of factors unrelated to attitudes*) to 9 (*not at all a result of factors unrelated to attitudes*). At end of the experiment, participants were debriefed.

Results

Table 12 displays descriptive information, Figure 7 shows a graphical representation of perceived validity means. Negative scores in the IAT represent a racial preference for Whites over Blacks.

We checked the number of participants who could not receive a score that deviated from their expectation because their expectation corresponded to one of the ends of the scale. This was only a small minority (<5%) and did not affect the results.⁸ See Figure 8 for the distribution of frequencies of self or others' explicit attitude per condition.

A multiple linear regression was conducted using perceived validity of the IAT as the dependent variable (the internal consistency of the scale was acceptable, $\alpha = .75$), and source of IAT result (self = -0.5, others = 0.5), desirability of deviation (desirable deviation = 0.5, undesirable deviation = -0.5), implicit attitude and explicit attitude for self or others as predictors. To

control for the impact of implicit and explicit attitude of self and others in our results, these two variables were centered, and we created interactions between each of these individual differences and our experimental manipulations, for a total of 11 parameters.

The regression model was significant, $F(11, 172) = 4.07, p < .001$, adjusted $R^2 = .16$. The effect of desirability of deviation was significant, $b = 0.68, SE = 0.27, 95\% CI [0.16, 1.20], p = .011$, as was the interaction between desirability of deviation and source, $b = -1.33, SE = 0.53, CI [-2.38, -0.29], p = .013$. The only other effect that was significant (see Table 13 for all regression coefficients) was a three-way interaction between desirability of deviation, source of IAT result, and explicit attitude of self or others, $b = -0.81, SE = 0.27, CI [-1.34, -0.28], p = .003$, such that as participants reported more positive attitudes for self or others (depending on condition), participants in the self/negative deviation condition (-0.5, -0.5) and in the others/positive deviation condition (coded as 0.5, 0.5) found the IAT to be less valid than those in the self/positive deviation condition (-0.5, 0.5) and those in the others/negative deviation condition (0.5, -0.5).

Given the significant interaction between desirability and source, we then proceeded to test self-other differences in the desirable and undesirable conditions. When the IAT result deviated in the negative attitude direction, participants in the self condition ($M = 3.32, SD = 2.06$) perceived the IAT to be less valid than did those in the others condition ($M = 4.42, SD = 1.48$), $t(74.94) = -2.92, p = .005, d = -0.61, 95\% CI [-1.02, -0.19]$. When the IAT result deviated in the positive attitude direction, results showed the opposite trend (self: $M = 5.18, SD = 1.59$ vs. others: $M = 4.66, SD = 1.22$), $t(89) = 1.72, p = .090, d = 0.36, CI [-0.06, 0.77]$.

Discussion

Experiment 6 demonstrated that the self-other asymmetry is not simply a result of confirming or disconfirming one's expectations. Using a manipulation that was individually tailored to reflect a

⁸ The pattern of results of this regression model holds even if we exclude from the analysis those few participants who could not receive a score that deviated two points from their expectation because their expectation was already at one of the ends of the scale or near it.

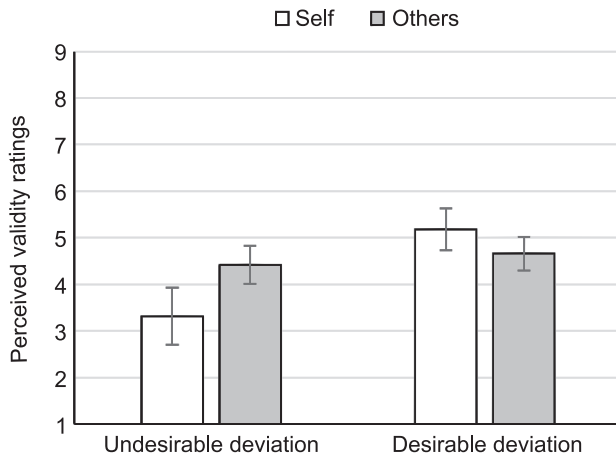


Figure 7. Mean Implicit Association Test perceived validity (with 95% confidence intervals), per condition (Experiment 6).

more negative or positive attitude than expected, we found that the IAT’s validity nevertheless fell along similar lines as in the prior experiments. That is, when it revealed a more desirable attitude than expected, it was deemed valid for the self but less valid for others, whereas when it revealed a more undesirable attitude than expected, it was deemed invalid for the self but more valid for others. These results garnered support for our hypothesis that self-enhancement motivation is an important factor in the way people make sense of IAT results.

Meta-Analysis of Individual Differences

Given that most of our experiments had only adequate power to find medium and large effects (see Table 14), we tested the impact of individual differences on perceived validity using internal meta-analyses. For each individual difference measured in more than one experiment, we computed two meta-analyses: (a) one for a three-way interaction among IAT result valence (positive or negative), source of IAT result (self or others), and the individual difference, and (b) one for the interaction between individual difference and IAT result valence. While the former addresses the question of whether the self-enhancement effect is moderated by individual differences, the latter tests whether individual differ-

ences exert an impact on perceived validity, depending only on the valence of the IAT result, regardless of the source of the result.

To obtain comparable standardized coefficients, we ensured that all partial effects included in the analysis were the result of regression models with the same number and type of parameters. Thus, we excluded Experiment 1 from the analysis, as there was no manipulation of valence of IAT result as a factor, and split Experiment 4 into two parts, one with Blacks as the attitudinal target and another with racists as the attitudinal target. In terms of software, we used the R metafor package (Viechtbauer, 2010) to calculate the meta-analytical effects, with standardized coefficient variances calculated using Aloe and Thompson’s (2013) equation (7). We present both fixed and random effects, as suggested by Goh, Hall, and Rosenthal (2016), and in line with the lack of precision that homogeneity tests reveal in small-scale meta-analyses (e.g., Bender et al., 2018). For random effects, we used restricted maximum-likelihood estimator and the Knapp-Hartung adjustment due to the small number of studies (see Bender et al., 2018).

The data used to calculate the meta-analyses can be found in Table 15 (for the implicit attitudes) and Table 16 (for explicit attitudes and implicit attitude awareness).

First, and most important for the current theorizing, do individual differences qualify the self-other asymmetry findings? Results suggest that that is not the case (see Figure 9). Implicit attitudes did not significantly interact with both valence and source of IAT result ($b = 0.03$, 95% CI $[-0.04, 0.09]$ for fixed, and $b = 0.03$, CI $[-0.02, 0.07]$ for random effects). Nor did explicit attitudes ($b = -0.09$, CI $[-0.20, 0.03]$ for fixed, and $b = -0.17$, CI $[-0.87, 0.53]$ for random effects), nor implicit attitude awareness ($b = -0.01$, CI $[-0.11, 0.09]$ for fixed, and $b = -0.01$, CI $[-0.19, 0.17]$ for random effects).

Did individual differences interact with valence of IAT result? The answer is, largely, yes (see Figure 10). There was a significant interaction of valence of IAT with impact of implicit attitudes ($b = 0.08$, CI $[0.00, 0.15]$ for fixed, and $b = 0.08$, CI $[0.02, 0.13]$ for random effects). Explicit attitude had a significant impact in the fixed effects, but not in the random effect analysis ($b = 0.32$, CI $[0.15, 0.48]$ for fixed, and $b = 0.39$, CI $[-0.31, 1.08]$ for random effects). Implicit attitude awareness had a significant effect in both analyses ($b = 0.27$, CI $[0.16, 0.39]$ for fixed effects, and $b = 0.27$, CI $[0.09, 0.46]$ for random effects).



Figure 8. Frequency of answers, per condition, on the Explicit Attitude Scale, for self or others. The labels of the explicit attitude ranged from 1 (extremely negative) to 9 (extremely positive).

Table 13
Regression Coefficients for Perceived Validity of the Implicit Association Test (IAT), Experiment 6

Parameter	B	SE	95% CI	β	p
Constant	4.41	.13	[4.15, 4.68]		<.001
Implicit attitude	-.18	.28	[-.72, .37]	-.04	.526
Explicit attitude (Self/Others)	-.01	.07	[-.15, .12]	-.02	.840
Source of IAT result	.25	.27	[-.28, .77]	.07	.354
Valence of deviation	.68	.27	[.16, 1.20]	.20	.011
Source \times Valence	-1.33	.53	[-2.38, -.29]	-.19	.013
Source \times Implicit	.18	.55	[-.91, 1.27]	.02	.743
Source \times Explicit (S/O)	.02	.13	[-.24, .29]	.01	.857
Valence \times Implicit	.11	.55	[-.97, 1.20]	.01	.837
Valence \times Explicit (S/O)	.15	.13	[-.11, .42]	.09	.263
Source \times Valence \times Implicit	.62	1.10	[-1.56, 2.79]	.04	.576
Source \times Valence \times Explicit (S/O)	-.81	.27	[-1.34, -.28]	-.24	.003

Note. CI = confidence interval.

In conclusion, for the individual differences we could meta-analyze, none reliably interacted with both valence and source of IAT result, but all had an impact in perceived validity that interacted with valence of IAT result, such that, when the IAT revealed a positive attitude, more positive implicit attitudes, explicit attitudes, or implicit attitude awareness led to higher perceived validity ratings. Yet, although these individual differences appear to have an important impact in perceptions of IAT validity regardless of who is the source of the IAT result (i.e., self or others), we did not find them to be significant moderators of the critical self-other asymmetry.

General Discussion

Six experiments demonstrated a systematic difference in the way people perceive the IAT's validity when its results refer to themselves as opposed to when they apply to other people, with the desirability of the IAT's result as a key moderator.

In the first experiment, participants who received an IAT result suggesting that they were prejudiced found the IAT to be less valid than did participants who evaluated the exact same result but this time referring to other people's prejudice. The second experiment replicated this finding, and it further showed the opposite pattern for a low-prejudice result: participants now believed the IAT to be more valid for themselves than for other people. The third experiment replicated those results with a different sample, method and target group (elderly people). The

fourth experiment demonstrated that this pattern of results can be reversed if the IAT involves targets for whom discrimination is prescribed (racists). The fifth experiment showed that the impact of source of IAT result and desirability spills over from the perception of how well the IAT measures attitudes to how adequate it is to use the IAT as a personnel selection tool for a job that involves interacting with Blacks. The final, sixth, experiment showed that a self-verification interpretation, whereby participants expect others to be more biased than themselves and therefore perceive the validity of the IAT as a function of these expectations, was not borne out. Indeed, when the result of the IAT was presented as a deviation from what participants reported as being their own, or others', attitude, the self-other asymmetry was still observed.

These results build on previous studies showing defensive reactions toward IAT scores (e.g., Hillard et al., 2013; Howell et al., 2013, 2015; Perry et al., 2015) and extend their implications by investigating how people perceive the IAT's validity with implications for its use in applied settings. In line with a motivational account, when the IAT applies to the self, people regard it as valid or invalid as a function of whether its results are desirable or not. When considering the same results for other people, however, there is no longer a motivation to think well of the test-takers, and so people may judge an undesirable result to be valid. Although in our experiments the other was always described as "most other people," we expect that when this other is, instead, someone who is close to us or who we particularly like, the pattern displayed will be closer to the self than to the others of our studies (Mata et al., 2018; Pedregon, Farley, Davis, Wood, & Clark, 2012).

Self-Enhancement or Other-Derogation?

A fundamental question at this point is whether the self-other differences observed in these studies are driven by self-enhancement or by other-derogation. Several rationales support interpreting the differences in terms of the desire to see oneself in a positive light rather than seeing others in a negative light. First, the studies used between-subjects designs, and therefore people made judgments only about the self or others, never both. Thus, people were not directly evaluating themselves as better than other

Table 14
Sensitivity Analysis for 80% Power for the Full Set of Experiments

Experiment	Sample size	Parameters	Sensitivity (f^2)
Experiment 1	52	5	.28
Experiment 2	74	7	.22
Experiment 3	87	15	.26
Experiment 4	190	31	.15
Experiment 5	111	7	.14
Experiment 6	184	11	.10

Note. Cutoff criteria for f^2 : small = .02, medium = .15, large = .35.

Table 15
Regression Data Used for the Meta-Analysis of Implicit Attitude

Variable	Implicit attitude					
	Experiment 2	Experiment 3	Experiment 4 (Blacks)	Experiment 4 (racists)	Experiment 5	Experiment 6
<i>n</i>	74	87	95	95	111	184
	Interaction with valence of IAT result (3 parameters)					
<i>R</i> ²	.141	.303	.228	.536	.262	.092
<i>b</i>	.040	.071	.092	.170	.065	.036
<i>r</i> _{sp}	.040	.069	.082	.135	.062	.036
	Interaction with valence and source of IAT result (7 parameters)					
<i>R</i> ²	.411	.514	.410	.631	.408	.155
<i>b</i>	.047	.113	-.016	-.008	.021	.009
<i>r</i> _{sp}	.044	.103	-.014	-.006	.020	.009

Note. IAT = Implicit Association Test. The table does not include Experiment 1 as it did not manipulate desirability of IAT result and to guarantee the same number of parameters per regression model, Experiment 4 is split by IAT target.

people, or others as worse than themselves. In conditions where the source of result was the self, participants reacted favorably to positive feedback and they dismissed negative feedback as invalid. When the IAT scores pertained to others, however, reactions were equivalent across the desirable versus undesirable conditions.

Indeed, an internal meta-analysis using McShane and Böckenholt's (2017) Shiny app revealed that the undesirable versus desirable contrast in the others condition is not reliably significant. We performed an internal meta-analysis with a 2 (source of IAT result: self vs. others) × 2 (desirability of IAT result: desirable vs. undesirable) design. In Experiment 1 we only included the IAT conditions (in which the result was of high prejudice, i.e., undesirable). Experiment 4 was split in two, such that the positive attitude toward racists condition was considered undesirable and the negative attitude toward racists was considered desirable. This resulted in a total of 7 experiments for the meta-analysis. First, this analysis showed that the self-other asymmetry is robust, both in the undesirable condition (estimate = -1.31, SE = 0.33) and in the desirable condition (estimate = 1.71, SE = 0.36). Of greater relevance to the present discussion, while the effect of desirability is significant in the self condition (estimate = 3.44, SE = 0.35), that is not the case in the others condition (estimate = 0.42, SE = 0.34). Figure 11 includes a visual representation of the meta-analytical effects and their respective 95% CIs.

Thus, what might look like a harsh judgment of others is the result of examining judgments for others in comparison to judgments about the self. Validity ratings were lower for others in the positive IAT result condition, but only when compared to the same ratings for the self, not when compared to ratings for others in the negative condition. Therefore, the interaction pattern is driven in large part by the self-referent ratings.

That does not mean, however, that these results are inconsequential for how people think about others. Experiment 5 revealed that a negative IAT result for the self led to lower IAT validity perceptions, which then led to higher dismissal toward the use of the IAT as a personnel selection tool, whereas the same result for others led to higher IAT validity perceptions, which then led to higher leniency toward the use of the IAT as a personnel selection tool for jobs where racial bias might be a concern. Thus, although the major differences in accepting IAT results are observed for the self, the consequences for judging others are of great relevance.

Individual Differences

We explored several individual differences that could influence the perceived validity of the IAT: (a) explicit attitudes, (b) implicit attitudes, (c) awareness of implicit attitudes, and (d) internal and external motivation to control prejudice. We first wish to note that,

Table 16
Regression Data Used for the Meta-Analysis of Explicit Attitude (Left Portion of the Table) and Implicit Attitude Awareness (Right Portion of the Table)

Variable	Explicit attitude			Implicit attitude awareness		
	Experiment 3	Experiment 4 (Blacks)	Experiment 4 (racists)	Experiment 3	Experiment 4 (Blacks)	Experiment 4 (racists)
<i>n</i>	87	95	95	87	95	95
	Interaction with valence of IAT result (3 parameters)					
<i>R</i> ²	.306	.330	.572	.352	.227	.570
<i>b</i>	.135	.652	.425	.205	.366	.287
<i>r</i> _{sp}	.114	.360	.109	.192	.276	.216
	Interaction with valence and source of IAT result (7 parameters)					
<i>R</i> ²	.536	.514	.693	.563	.467	.680
<i>b</i>	-.149	.001	-.648	.062	-.020	-.072
<i>r</i> _{sp}	-.124	.001	-.144	.056	-.015	-.052

Note. IAT = Implicit Association Test.

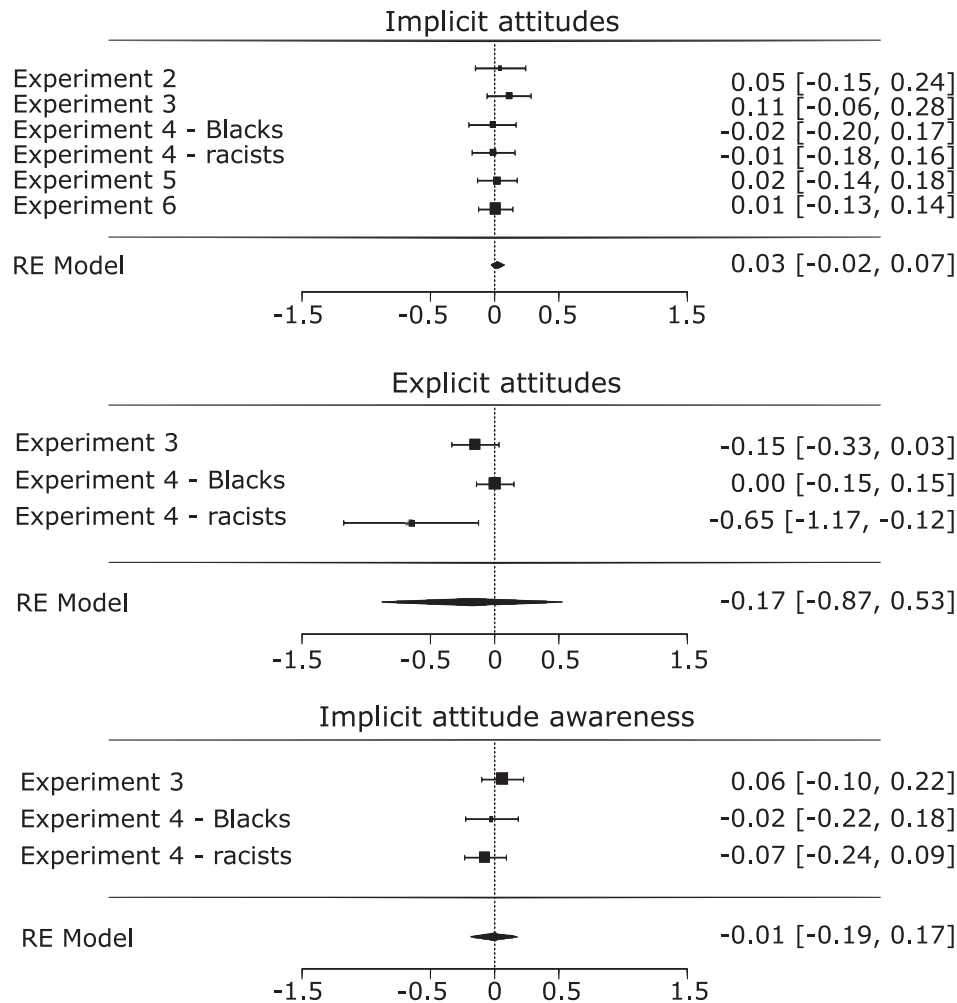


Figure 9. Forest plots of the random effects internal meta-analysis of the regression beta coefficients for the three-way interaction between valence of Implicit Association Test (IAT) result, source of IAT result (self vs. others), and each of the three individual differences: implicit attitudes, explicit attitudes, and implicit attitude awareness.

although our experiments had adequate power to detect the effects relating to the critical self-other asymmetry, which we expected to be medium to large in line with the large self-other asymmetries found in the literature (see the meta-analysis by Heine & Hamamura, 2007; and Study 3 of Hahn et al., 2014, which shows a self-other asymmetry in implicit bias predictions), most of the regression analyses could only detect medium to large effects (see Table 14). Hence, conclusions regarding these individual factors are limited by modest sample sizes.

A meta-analytical approach enabled us to test the impact of these individual differences. It found that explicit attitudes, implicit attitudes, and awareness of implicit attitudes all had a significant, positive impact of perceived validity when the IAT revealed a positive attitude. Of these three, awareness of implicit attitudes had the strongest effect, but confidence was greater in the estimation of the impact of implicit attitudes, as these were measured in six studies, totaling 646 participants, whereas explicit

attitudes and awareness of implicit attitudes were only measured in three experiments, totaling 277 participants.

Turning to the individual differences not included in the meta-analysis, internal and external motivation to control prejudice, no significant effect was found in the one experiment (Experiment 4) that explored their relationship with perceived validity, albeit with low power. Still, these scales may have an important role in making sense of IAT validity perceptions, as they may tap into what participants find personally (i.e., internally) or socially (i.e., externally) desirable. Because we found desirability to be a main component of how people make sense of an IAT result, what people find internally or externally desirable should have an impact on perceived validity of the IAT result. Again, larger samples may be needed to find the impact of these motivations, or, alternatively, one might measure what each person finds desirable in other ways (e.g., by directly asking participants what result they consider the most desirable).

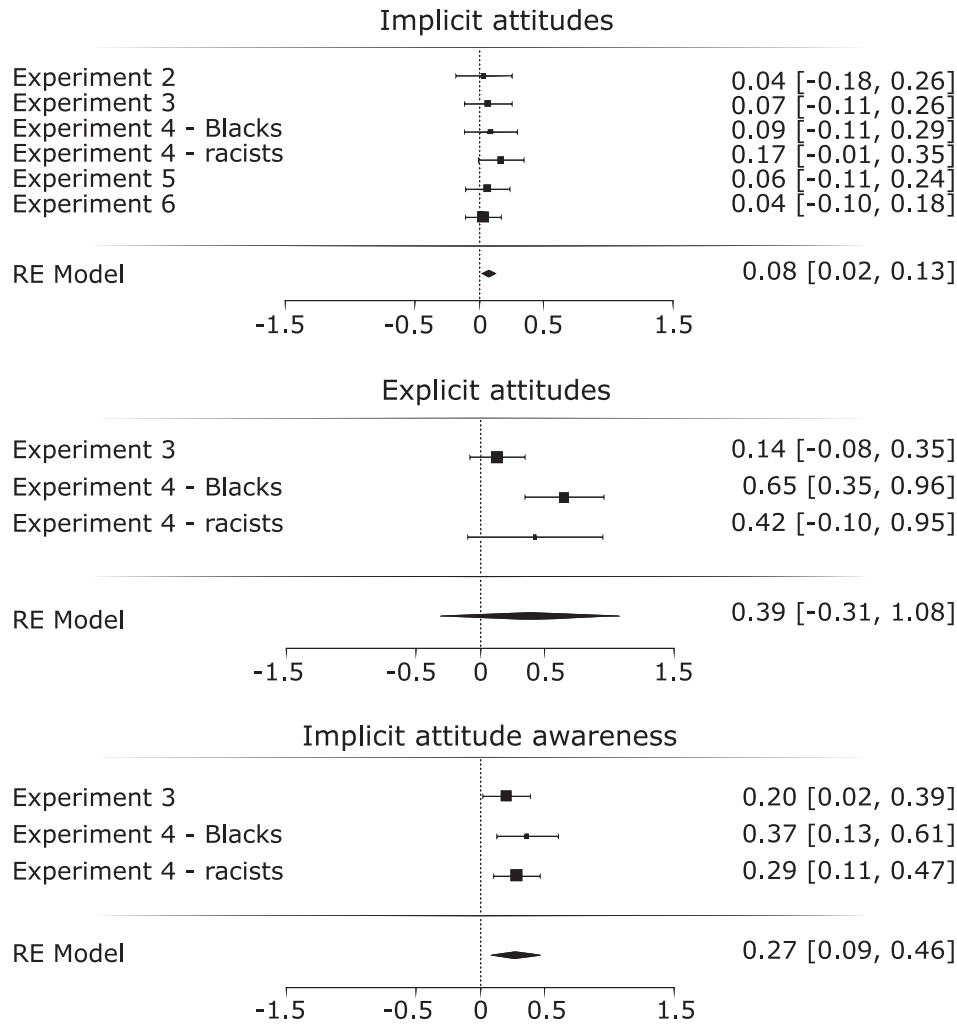


Figure 10. Forest plots of the random effects internal meta-analysis of the regression beta coefficients for the two-way interaction between valence of Implicit Association Test (IAT) result and each of the three individual differences: implicit attitudes, explicit attitudes, and implicit attitude awareness.

Limitations and Future Studies

Throughout six studies, we found important situational and individual-difference factors that have an impact on the perceived validity of the IAT when the test results refer to one own's performance on the IAT, but only individual differences seem to have a reliable impact on reactions to the IAT when its results refer to others. Although this may be attributed to the low power of the present studies to detect small effects (see Table 14; though this is in part addressed by the meta-analyses presented above), it may also be the case that the situational factors that influence people's reactions to the IAT are different when the results refer to others versus themselves. Future studies might address this question, and explore potential characteristics of others (e.g., being a close other or a stranger) that might lead to IAT derogation versus acceptance.

The current experiments did not investigate how members of minority groups judge the validity of the IAT. This was due to the small size of the minorities subsample in our experiments (Black

participants: 3.85–11.67%; biracial: 2.07–5.13%). Howell et al. (2015) found that, although both White and Black individuals were defensive when receiving feedback that deviated from their explicit attitudes in the pro-White direction, biracial Black/White individuals reacted defensively to deviations from their expectation in either direction. Howell et al. (2015) suggested that these results are a consequence of what the different groups value (to avoid appearing racist, in-group favoritism, and egalitarianism) and we agree with this intuition as it is in line with our desirability interpretation. Additional research attempting to disentangle desirability from explicit attitudes, for example, using a paradigm similar to Experiment 6's (in which IAT result was determined as a function of explicit attitude), could be potentially fruitful especially for validity perceptions among minority and multiple-identity individuals.

Moreover, the current set of studies used a between-subjects design, such that participants reacted to IAT results referring to them-

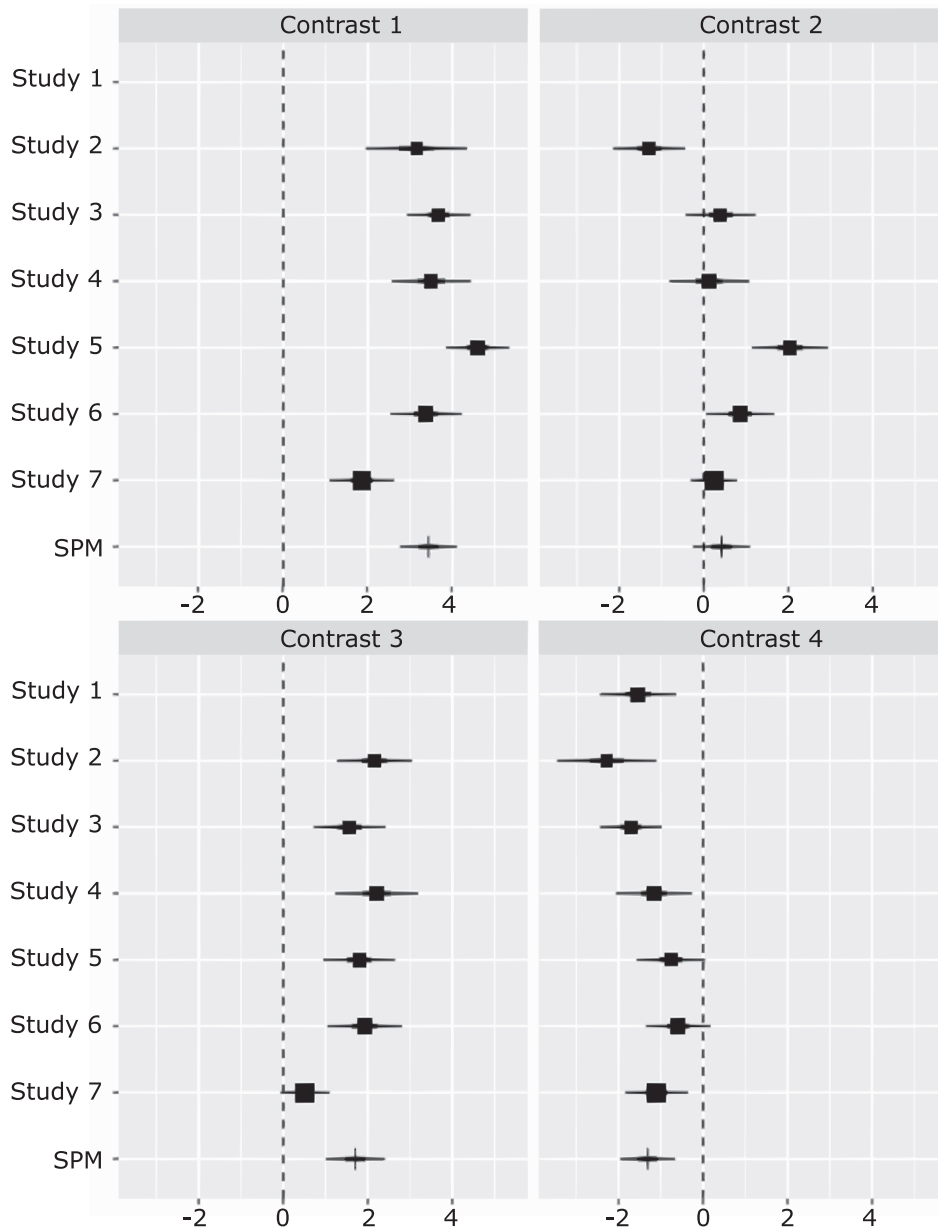


Figure 11. Plot of the single paper meta-analysis with all possible contrasts between desirability of Implicit Association Test (IAT) result and source of IAT result. Thick bars represent 50% confidence interval (CI), while thinner bars represent 95% CI. SPM = the single paper meta-analytical effect; contrast 1 = difference between desirable and undesirable results within the self condition; contrast 2 = difference between desirable and undesirable results within the others condition; contrast 3 = difference between self and others in the desirable condition; contrast 4 = difference between self and others within the undesirable condition. Note that Study 4 was split into two (Study 4 = Blacks condition and Study 5 = racists condition of Experiment 4).

This document is copyrighted by the American Psychological Association or one of its allied publishers. This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

selves or others, never both. This raises the question of whether the same person might have different reactions to the same test for self versus others. To test for that would call for a within-subjects design, which may lead to plausibility constraints (Kunda, 1990). In particular, we suspect that participants would catch themselves defending a double standard and might be constrained to defend that the IAT says something about themselves but not about others, or vice versa. A

study with such a within-subjects design might make for a strong test of how strategic and malleable people are in shifting their reactions to a test depending on the desirability of its results (Mata et al., 2013), but would be less representative of more common situations where people either take the test or they hear of another person's test result, not both. Still, this is an intriguing possibility that might be investigated in future research.

Implications

The self-other asymmetry found in our experiments should caution intentions to use the IAT in applied settings, as doing so may lead to disagreement between those who administer the test, or at least take decisions on the basis of its results (e.g., employers, judges/juries, teachers) and those who take the test (e.g., employees, defendants, students), and who may therefore react defensively toward their evaluation. These disagreements may consist of a lack of trust on those endorsing the IAT, perceptions of injustice, and ineffective behavior change, in line with research on the impact of persuasiveness of source credibility (Pornpitakpan, 2004) and the impact of perceived validity in personnel selection processes (Smither et al., 1993).

Furthermore, our experiments present challenges to the suggestions of providing IAT feedback with the aim of raising implicit bias awareness. If a significant proportion of people who receive undesirable IAT feedback think of the IAT as being an invalid measure of attitudes, they may gain little by going through such an experience. This idea is supported by Hillard et al.'s (2013) study where increased guilt and helpful behaviors toward Blacks were observed in only a small minority (12%) of participants—those who did not reject the notion that the feedback reflected personal attitudes despite getting feedback that they were biased. This finding suggests that accepting that the IAT reflects something meaningful about one's attitudes may be an important ingredient in benefiting from IAT feedback.

Conclusion

When predicting whether people perceive the IAT to be a valid instrument, it is important to consider both its results and to whom they refer. People display defensiveness when the test is said to reveal that they harbor undesirable prejudices and endorse the test when it makes them out to be not prejudicial. When the test results pertain to other people's bias, its validity is more or less unaffected by what it says about their degree of bias. This self-other asymmetry may lead to differences in how people first react to the IAT and then to its applications in specific contexts. Thus, when providing feedback to the IAT or considering its use in applied settings, it is important to consider the reactions of both test-takers and of those who use the IAT's results for judgment and decision making.

References

- Alicke, M. D. (1985). Global self-evaluation as determined by the desirability and controllability of trait adjectives. *Journal of Personality and Social Psychology, 49*, 1621–1630. <http://dx.doi.org/10.1037/0022-3514.49.6.1621>
- Aloe, A. M., & Thompson, C. G. (2013). The synthesis of partial effect sizes. *Journal of the Society for Social Work and Research, 4*, 390–405. <http://dx.doi.org/10.5243/jsswr.2013.24>
- Ayres, I. (2001). *Pervasive prejudice? Unconventional evidence of race and gender discrimination*. Illinois: The University of Chicago Press.
- Bar-Anan, Y., & Nosek, B. A. (2014). A comparative investigation of seven indirect attitude measures. *Behavior Research Methods, 46*, 668–688. <http://dx.doi.org/10.3758/s13428-013-0410-6>
- Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad is stronger than good. *Review of General Psychology, 5*, 323–370. <http://dx.doi.org/10.1037/1089-2680.5.4.323>
- Bender, R., Friede, T., Koch, A., Kuss, O., Schlattmann, P., Schwarzer, G., & Skipka, G. (2018). Methods for evidence synthesis in the case of very few studies. *Research Synthesis Methods, 9*, 382–392. <http://dx.doi.org/10.1002/jrsm.1297>
- Bendick, M., Jr., & Nunes, A. P. (2012). Developing the research basis for controlling bias in hiring. *Journal of Social Issues, 68*, 238–262. <http://dx.doi.org/10.1111/j.1540-4560.2012.01747.x>
- Blanton, H., Jaccard, J., Gonzales, P. M., & Christie, C. (2006). Decoding the Implicit Association Test: Implications for criterion prediction. *Journal of Experimental Social Psychology, 42*, 192–212. <http://dx.doi.org/10.1016/j.jesp.2005.07.003>
- Bluemke, M., & Friese, M. (2008). Reliability and validity of the Single-Target IAT (ST-IAT): Assessing automatic affect towards multiple attitude objects. *European Journal of Social Psychology, 38*, 977–997. <http://dx.doi.org/10.1002/ejsp.487>
- Brendl, C. M., Markman, A. B., & Messner, C. (2001). How do indirect measures of evaluation work? Evaluating the inference of prejudice in the Implicit Association Test. *Journal of Personality and Social Psychology, 81*, 760–773. <http://dx.doi.org/10.1037/0022-3514.81.5.760>
- Brigham, J. C. (1993). College students' racial attitudes. *Journal of Applied Social Psychology, 23*, 1933–1967. <http://dx.doi.org/10.1111/j.1559-1816.1993.tb01074.x>
- Brown, J. D. (2012). Understanding the better than average effect: Motives (still) matter. *Personality and Social Psychology Bulletin, 38*, 209–219. <http://dx.doi.org/10.1177/0146167211432763>
- Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science, 6*, 3–5. <http://dx.doi.org/10.1177/1745691610393980>
- Carpenter, T. P., Pogacar, R., Pullig, C., Kouril, M., Aguilar, S., LaBouff, J., . . . Chakroff, A. (2018). *Conducting IAT research within online surveys: A procedure, validation, and open source tool*. <http://dx.doi.org/10.17605/OSF.IO/6XDYJ>
- Chambers, J. R., & Windschitl, P. D. (2004). Biases in social comparative judgments: The role of nonmotivated factors in above-average and comparative-optimism effects. *Psychological Bulletin, 130*, 813–838. <http://dx.doi.org/10.1037/0033-2909.130.5.813>
- Chan, D. (1997). Racial subgroup differences in predictive validity perceptions on personality and cognitive ability tests. *Journal of Applied Psychology, 82*, 311–320. <http://dx.doi.org/10.1037/0021-9010.82.2.311>
- Chan, D., Schmitt, N., DeShon, R. P., Clause, C. S., & Delbridge, K. (1997). Reactions to cognitive ability tests: The relationships between race, test performance, face validity perceptions, and test-taking motivation. *Journal of Applied Psychology, 82*, 300–310. <http://dx.doi.org/10.1037/0021-9010.82.2.300>
- Chan, D., Schmitt, N., Jennings, D., Clause, C. S., & Delbridge, K. (1998). Applicant perceptions of test fairness: Integrating justice and self-serving bias perspectives. *International Journal of Selection and Assessment, 6*, 232–239. <http://dx.doi.org/10.1111/1468-2389.00094>
- Chan, D., Schmitt, N., Sacco, J. M., & DeShon, R. P. (1998). Understanding pretest and posttest reactions to cognitive ability and personality tests. *Journal of Applied Psychology, 83*, 471–485. <http://dx.doi.org/10.1037/0021-9010.83.3.471>
- Chandler, J., Mueller, P., & Paolacci, G. (2014). Nonnaïveté among Amazon Mechanical Turk workers: Consequences and solutions for behavioral researchers. *Behavior Research Methods, 46*, 112–130. <http://dx.doi.org/10.3758/s13428-013-0365-7>
- Codol, J.-P. (1975). On the so-called "superior conformity of the self" behavior: Twenty experimental investigations. *European Journal of Social Psychology, 5*, 457–501. <http://dx.doi.org/10.1002/ejsp.2420050404>
- Crandall, C. S., Eshleman, A., & O'Brien, L. (2002). Social norms and the expression and suppression of prejudice: The struggle for internalization.

- Journal of Personality and Social Psychology*, 82, 359–378. <http://dx.doi.org/10.1037/0022-3514.82.3.359>
- 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), e57410. <http://dx.doi.org/10.1371/journal.pone.0057410>
- Devine, P. G., Plant, E. A., Amodio, D. M., Harmon-Jones, E., & Vance, S. L. (2002). The regulation of explicit and implicit race bias: The role of motivations to respond without prejudice. *Journal of Personality and Social Psychology*, 82, 835–848. <http://dx.doi.org/10.1037/0022-3514.82.5.835>
- Ditto, P. H., & Lopez, D. F. (1992). Motivated skepticism: Use of differential decision criteria for preferred and nonpreferred conclusions. *Journal of Personality and Social Psychology*, 63, 568–584. <http://dx.doi.org/10.1037/0022-3514.63.4.568>
- Ditto, P. H., Munro, G. D., Apanovitch, A. M., Scepansky, J. A., & Lockhart, L. K. (2003). Spontaneous skepticism: The interplay of motivation and expectation in responses to favorable and unfavorable medical diagnoses. *Personality and Social Psychology Bulletin*, 29, 1120–1132. <http://dx.doi.org/10.1177/0146167203254536>
- Dovidio, J. F., Kawakami, K., & Gaertner, S. L. (2002). Implicit and explicit prejudice and interracial interaction. *Journal of Personality and Social Psychology*, 82, 62–68. <http://dx.doi.org/10.1037/0022-3514.82.1.62>
- Ebner, N. C., Riediger, M., & Lindenberger, U. (2010). FACES—a database of facial expressions in young, middle-aged, and older women and men: Development and validation. *Behavior Research Methods*, 42, 351–362. <http://dx.doi.org/10.3758/BRM.42.1.351>
- Erdelyi, M. H. (1974). A new look at the new look: Perceptual defense and vigilance. *Psychological Review*, 81, 1–25. <http://dx.doi.org/10.1037/h0035852>
- Fiedler, K., Messner, C., & Bluemke, M. (2006). Unresolved problems with the “I”, the “A”, and the “T”: A logical and psychometric critique of the Implicit Association Test (IAT). *European Review of Social Psychology*, 17, 74–147. <http://dx.doi.org/10.1080/10463280600681248>
- Friese, M., Bluemke, M., & Wänke, M. (2007). Predicting voting behavior with implicit attitude measures: The 2002 German parliamentary election. *Experimental Psychology*, 54, 247–255. <http://dx.doi.org/10.1027/1618-3169.54.4.247>
- Friese, M., Smith, C. T., Plischke, T., Bluemke, M., & Nosek, B. A. (2012). Do implicit attitudes predict actual voting behavior particularly for undecided voters? *PLoS ONE*, 7(8), e44130. <http://dx.doi.org/10.1371/journal.pone.0044130>
- Gibson, B. (2008). Can evaluative conditioning change attitudes toward mature brands? New evidence from the Implicit Association Test. *Journal of Consumer Research*, 35, 178–188. <http://dx.doi.org/10.1086/527341>
- Goh, J. X., Hall, J. A., & Rosenthal, R. (2016). Mini meta-analysis of your own studies: Some arguments on why and a primer on how. *Social and Personality Psychology Compass*, 10, 535–549. <http://dx.doi.org/10.1111/spc3.12267>
- Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review*, 102, 4–27. <http://dx.doi.org/10.1037/0033-295X.102.1.4>
- Greenwald, A. G., Banaji, M. R., Rudman, L. A., Farnham, S. D., Nosek, B. A., & Mellott, D. S. (2002). A unified theory of implicit attitudes, stereotypes, self-esteem, and self-concept. *Psychological Review*, 109, 3–25. <http://dx.doi.org/10.1037/0033-295X.109.1.3>
- Greenwald, A. G., & Krieger, L. H. (2006). Implicit bias: Scientific foundations. *California Law Review*, 94, 945–967. <http://dx.doi.org/10.2307/20439056>
- 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, 1464–1480. <http://dx.doi.org/10.1037/0022-3514.74.6.1464>
- Greenwald, A. G., Nosek, B. A., & Banaji, M. R. (2003). Understanding and using the implicit association test: I. An improved scoring algorithm. *Journal of Personality and Social Psychology*, 85, 197–216. <http://dx.doi.org/10.1037/0022-3514.85.2.197>
- Hahn, A., Judd, C. M., Hirsh, H. K., & Blair, I. V. (2014). Awareness of implicit attitudes. *Journal of Experimental Psychology: General*, 143, 1369–1392. <http://dx.doi.org/10.1037/a0035028>
- Handelsman, J., & Sakraney, N. (2015). *Implicit bias* [PDF document]. Retrieved from https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/bias_9-14-15_final.pdf
- Hayes, A. F. (2013). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach*. New York, NY: Guilford Press.
- Heine, S. J., & Hamamura, T. (2007). In search of East Asian self-enhancement. *Personality and Social Psychology Review*, 11, 4–27. <http://dx.doi.org/10.1177/1088868306294587>
- Hillard, A. L., Ryan, C. S., & Gervais, S. J. (2013). Reactions to the Implicit Association Test as an educational tool: A mixed methods study. *Social Psychology of Education: An International Journal*, 16, 495–516. <http://dx.doi.org/10.1007/s11218-013-9219-5>
- Howell, J. L., Collisson, B., Crysel, L., Garrido, C. O., Newell, S. M., Cottrell, C. A., . . . Shepperd, J. A. (2013). Managing the threat of impending implicit attitude feedback. *Social Psychological and Personality Science*, 4, 714–720. <http://dx.doi.org/10.1177/1948550613479803>
- Howell, J. L., Gaither, S. E., & Ratliff, K. A. (2015). Caught in the middle: Defensive responses to IAT feedback among Whites, Blacks, and biracial Black/Whites. *Social Psychological and Personality Science*, 6, 373–381. <http://dx.doi.org/10.1177/1948550614561127>
- Kang, J. (2005). Trojan horses of race. *Harvard Law Review*, 118, 1489–1593. Retrieved from <http://www.jstor.org/stable/4093447?origin=JSTOR-pdf>
- Kang, J., & Banaji, M. R. (2006). Fair measures: A behavioral realist revision of “affirmative action”. *California Law Review*, 94, 1063–1118. <http://dx.doi.org/10.2307/20439059>
- Kang, J., Bennett, M. W., Carbado, D. W., Casey, P., Dasgupta, N., Faigman, D. L., . . . Mnookin, J. L. (2012). Implicit bias in the courtroom. *UCLA Law Review*, 59, 1124–1186. Retrieved from <https://www.uclalawreview.org/implicit-bias-in-the-courtroom-2/>
- Klein, N., & Epley, N. (2016). Maybe holier, but definitely less evil, than you: Bounded self-righteousness in social judgment. *Journal of Personality and Social Psychology*, 110, 660–674. <http://dx.doi.org/10.1037/pspa0000050>
- Klein, N., & Epley, N. (2017). Less evil than you: Bounded self-righteousness in character inferences, emotional reactions, and behavioral extremes. *Personality and Social Psychology Bulletin*, 43, 1202–1212. <http://dx.doi.org/10.1177/0146167217711918>
- Krieger, L. H., & Fiske, S. T. (2006). Behavioral realism in employment discrimination law: Implicit bias and disparate treatment. *California Law Review*, 94, 997–1062. <http://dx.doi.org/10.2307/20439058>
- Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, 108, 480–498. <http://dx.doi.org/10.1037/0033-2909.108.3.480>
- Lang, P. J., Bradley, M. M., & Cuthbert, B. N. (2008). *International affective picture system (IAPS): Affective ratings of pictures and instruction manual* (Tech. Rep. No. A-8.) Gainesville: University of Florida.
- Lord, C. G., Ross, L., & Lepper, M. R. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology*, 37, 2098–2109. <http://dx.doi.org/10.1037/0022-3514.37.11.2098>
- Mata, A., Ferreira, M. B., & Sherman, S. J. (2013). Flexibility in motivated reasoning: Strategic shifts of reasoning modes in covariation judgment. *Social Cognition*, 31, 465–481. http://dx.doi.org/10.1521/soco_2012_1004

- Mata, A., Fiedler, K., Ferreira, M. B., & Almeida, T. (2013). Reasoning about others' reasoning. *Journal of Experimental Social Psychology, 49*, 486–491. <http://dx.doi.org/10.1016/j.jesp.2013.01.010>
- Mata, A., Sherman, S. J., Ferreira, M. B., & Mendonça, C. (2015). Strategic numeracy: Self-serving reasoning about health statistics. *Basic and Applied Social Psychology, 37*, 165–173. <http://dx.doi.org/10.1080/01973533.2015.1018991>
- Mata, A., Simão, C., Farias, A. R., & Steimer, A. (2018). Forecasting emotional duration: A motivational account and self-other differences. *Emotion*. Advance online publication. <http://dx.doi.org/10.1037/emo0000455>
- McShane, B. B., & Böckenholt, U. (2017). Single paper meta-analysis: Benefits for study summary, theory-testing, and replicability. *Journal of Consumer Research, 43*, 1048–1063. <http://dx.doi.org/10.1093/jcr/ucw085>
- Miller, D. T., & Ross, M. (1975). Self-serving biases in the attribution of causality: Fact or fiction? *Psychological Bulletin, 82*, 213–225. <http://dx.doi.org/10.1037/h0076486>
- Mitamura, C., Erickson, L., & Devine, P. G. (2017). Value-based standards guide sexism inferences for self and others. *Journal of Experimental Social Psychology, 72*, 101–117. <http://dx.doi.org/10.1016/j.jesp.2017.04.006>
- Monteith, M. J., Voils, C. I., & Ashburn-Nardo, L. (2001). Taking a look underground: Detecting, interpreting, and reacting to implicit racial biases. *Social Cognition, 19*, 395–417. <http://dx.doi.org/10.1521/soco.19.4.395.20759>
- Morris, K. A., & Ashburn-Nardo, L. (2010). The Implicit Association Test as a class assignment: Student affective and attitudinal reactions. *Teaching of Psychology, 37*, 63–68. <http://dx.doi.org/10.1080/00986280903426019>
- Morrison, M., DeVaul-Fetters, A., & Gawronski, B. (2016). Stacking the jury: Legal professionals' peremptory challenges reflect jurors' levels of implicit race bias. *Personality and Social Psychology Bulletin, 42*, 1129–1141. <http://dx.doi.org/10.1177/0146167216651853>
- Nisbett, R. E., & Ross, L. (1980). *Human inference: Strategies and shortcomings of social judgment*. Englewood Cliffs, NJ: Prentice Hall.
- Norton, M. I., Sommers, S. R., & Brauner, S. (2007). Bias in jury selection: Justifying prohibited peremptory challenges. *Journal of Behavioral Decision Making, 20*, 467–479. <http://dx.doi.org/10.1002/bdm.571>
- Nosek, B. A., Greenwald, A. G., & Banaji, M. R. (2005). Understanding and using the Implicit Association Test: II. Method variables and construct validity. *Personality and Social Psychology Bulletin, 31*, 166–180. <http://dx.doi.org/10.1177/0146167204271418>
- Nosek, B. A., & Hansen, J. J. (2008). The associations in our heads belong to us: Searching for attitudes and knowledge in implicit evaluation. *Cognition and Emotion, 22*, 553–594. <http://dx.doi.org/10.1080/02699930701438186>
- Nosek, B. A., Smyth, F. L., Hansen, J. J., Devos, T., Lindner, N. M., Ranganath, K. A., . . . Banaji, M. R. (2007). Pervasiveness and correlates of implicit attitudes and stereotypes. *European Review of Social Psychology, 18*, 36–88. <http://dx.doi.org/10.1080/10463280701489053>
- Ottaway, S. A., Hayden, D. C., & Oakes, M. A. (2001). Implicit attitudes and racism: Effects of word familiarity and frequency on the Implicit Association Test. *Social Cognition, 19*, 97–144. <http://dx.doi.org/10.1521/soco.19.2.97.20706>
- Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making, 5*, 411–419. Retrieved from <http://journal.sjdm.org/vol5.5.html>
- Pedregon, C. A., Farley, R. L., Davis, A., Wood, J. M., & Clark, R. D. (2012). Social desirability, personality questionnaires, and the “better than average” effect. *Personality and Individual Differences, 52*, 213–217. <http://dx.doi.org/10.1016/j.paid.2011.10.022>
- Peer, E., Vosgerau, J., & Acquisti, A. (2014). Reputation as a sufficient condition for data quality on Amazon Mechanical Turk. *Behavior Research Methods, 46*, 1023–1031. <http://dx.doi.org/10.3758/s13428-013-0434-y>
- Perry, S. P., Murphy, M. C., & Dovidio, J. F. (2015). Modern prejudice: Subtle, but unconscious? The role of bias awareness in White's perceptions of personal and others' bias. *Journal of Experimental Social Psychology, 61*, 64–78. <http://dx.doi.org/10.1016/j.jesp.2015.06.007>
- Petty, R. E., & Cacioppo, J. T. (1979). Issue-involvement can increase or decrease persuasion by enhancing message-relevant cognitive responses. *Journal of Personality and Social Psychology, 37*, 1915–1926. <http://dx.doi.org/10.1037/0022-3514.37.10.1915>
- Plant, E. A., & Devine, P. G. (1998). Internal and external motivation to respond without prejudice. *Journal of Personality and Social Psychology, 75*, 811–832. <http://dx.doi.org/10.1037/0022-3514.75.3.811>
- Pornpitakpan, C. (2004). The persuasiveness of source credibility: A critical review of five decades' evidence. *Journal of Applied Social Psychology, 34*, 243–281. <http://dx.doi.org/10.1111/j.1559-1816.2004.tb02547.x>
- Pronin, E., Lin, D. Y., & Ross, L. (2002). The bias blind spot: Perceptions of bias in self versus others. *Personality and Social Psychology Bulletin, 28*, 369–381. <http://dx.doi.org/10.1177/0146167202286008>
- Rachlinski, J. J., Johnson, S. L., Wistrich, A. J., & Guthrie, C. (2009). Does unconscious racial bias affect trial judges? *The Notre Dame Law Review, 84*, 1195–1246. Retrieved from <http://ndlawreview.org/publications/archives/volume-84/issue-3/>
- Rooth, D.-O. (2007). Evidence of unequal treatment in hiring against obese applicants: A field experiment. *IZA Discussion Paper Series, 2775*. Retrieved from http://legacy.iza.org/en/webcontent/publications/papers/viewAbstract?dp_id=2775
- Saujani, R. M. (2003). “The Implicit Association Test”: A measure of unconscious racism in legislative decision-making. *Michigan Journal of Race & Law, 8*, 395–423. Retrieved from <http://heinonline.org/HOL/LandingPage?handle=hein.journals/mjrl8&div=15>
- Smither, J. W., Reilly, R. R., Millsap, R. E., Pearlman, K., & Stoffey, R. W. (1993). Applicant reactions to selection procedures. *Personnel Psychology, 46*, 49–76. <http://dx.doi.org/10.1111/j.1744-6570.1993.tb00867.x>
- Staats, C., Capatosto, K., Wright, R. A., & Jackson, V. W. (2016). *State of the science: Implicit bias review 2016*. Columbus, OH: Kirwan Institute. Retrieved from <http://kirwaninstitute.osu.edu/my-product/2016-state-of-the-science-implicit-bias-review/>
- Steimer, A., & Mata, A. (2016). Motivated implicit theories of personality: My weaknesses will go away, but my strengths are here to stay. *Personality and Social Psychology Bulletin, 42*, 415–429. <http://dx.doi.org/10.1177/0146167216629437>
- Swann, W. B., Jr. (1983). Self-verification: Bringing social reality into harmony with the self. In J. Suls & A. G. Greenwald (Eds.), *Psychological perspectives on the self* (Vol. 2, pp. 33–66). Hillsdale, NJ: Erlbaum.
- Swann, W. B., Jr., Griffin, J. J., Jr., Predmore, S. C., & Gaines, B. (1987). The cognitive-affective crossfire: When self-consistency confronts self-enhancement. *Journal of Personality and Social Psychology, 52*, 881–889. <http://dx.doi.org/10.1037/0022-3514.52.5.881>
- Teachman, B. A., & Brownell, K. D. (2001). Implicit anti-fat bias among health professionals: Is anyone immune? *International Journal of Obesity, 25*, 1525–1531. <http://dx.doi.org/10.1038/sj.ijo.0801745>
- Tetlock, P. E., & Levi, A. (1982). Attribution bias: On the inconclusiveness of the cognition-motivation debate. *Journal of Experimental Social Psychology, 18*, 68–88. [http://dx.doi.org/10.1016/0022-1031\(82\)90082-8](http://dx.doi.org/10.1016/0022-1031(82)90082-8)
- Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software, 36*, 1–48. <http://www.jstatsoft.org/v36/i03/>. <http://dx.doi.org/10.18637/jss.v036.i03>
- Zhou, H., & Fishbach, A. (2016). The pitfall of experimenting on the web: How unattended selective attrition leads to surprising (yet false) research

- conclusions. *Journal of Personality and Social Psychology*, *111*, 493–504. <http://dx.doi.org/10.1037/pspa0000056>
- Ziegert, J. C., & Hanges, P. J. (2005). Employment discrimination: The role of implicit attitudes, motivation, and a climate for racial bias. *Journal of Applied Psychology*, *90*, 553–562. <http://dx.doi.org/10.1037/0021-9010.90.3.553>
- Zitek, E. M., & Hebl, M. R. (2007). The role of social norm clarity in the influenced expression of prejudice over time. *Journal of Experimental Social Psychology*, *43*, 867–876. <http://dx.doi.org/10.1016/j.jesp.2006.10.010>

Received July 28, 2017

Revision received November 6, 2018

Accepted December 21, 2018 ■