INTROSPECTION OF IMPLICIT ATTITUDES

By

ADAM HAHN

Diplom in Psychology, Freie Universität Berlin, 2007

M.A., Unversity of Colorado Boulder, 2009

Dissertation submitted to the Faculty of the Graduate School of the University of Colorado Boulder in partial fulfillment of the requirement for the degree of Doctor of Philosophy in Social

Department of Psychology and Neuroscience

2012

This thesis entitled:
Introspection of Implicit Attitudes
written by Adam Hahn
has been approved for the Department of Psychology and Neuroscience

Charles Judd

Irene Blair

Date May 10, 2012

The final copy of this thesis has been examined by the signatories, and we
Find that both the content and the form meet acceptable presentation standards
Of scholarly work in the above mentioned discipline.

IRB protocol # 0309.17

Hahn, Adam (Ph.D., Psychology and Neuroscience)

Introspection of Implicit Attitudes

Dissertation directed by professor Charles Judd

This dissertation addresses the general assumption that the implicit evaluative associations people might hold with social groups (i.e., usually referred to as "implicit attitudes") are "unconscious" and introspectively unavailable. In the work presented in the current dissertation I directly asked participants to predict their results on five future IATs. I consistently found that participants were highly accurate in their predictions, regardless of whether the IATs were described as revealing true attitudes or cultural associations (Studies 1 and 2); whether predictions were made in the form of specific response patterns ("ease of responding" in Study 1) or a more conceptual response ("your implicit attitude" in Studies 2-5); regardless of how much experience or explanation participants received before making their predictions (Study 4); and regardless of how much deliberation about the attitude targets participants engaged in prior to predicting their implicit attitudes (Study 5). Results also suggested that participants had unique insight into their own implicit attitudes above assumptions of normative responses, as their predictions were accurate even when controlling for predictions they made about the implicit attitude scores for a typical other participant in the study (Study 3). Even as participants accurately predicted their implicit attitudes, they reported distinct explicit attitudes. Interestingly, although participants showed impressive accuracy in predicting how their own implicit attitude scores would relate to each other, they showed somewhat limited insight into how their implicit attitudes compared to those of other people. These results fit theoretical dual-process models on attitudes, and they have several theoretical and practical implications.

DEDICATION

I would like to dedicate this dissertation to my friends and family in Germany, the United States, and elsewhere. Their patience, love, and support across oceans and continents made the completion of my doctorate possible. I feel enormous gratitude for the confidence and faith invested in me and my work throughout this journey.

ACKNOWLEDGEMENTS

I owe the completion of my doctoral work to the incredible support from my advisors. I especially wish to thank:

Chick Judd and Bernadette Park. Their undying support, belief in my ideas and abilities, willingness to listen and provide guidance in all matters (research or otherwise), but above all, their friendship, allowed for the necessary perseverance and patience to finish the requirements for a Ph.D., acclimate to life in a previously foreign culture, and become a true researcher and a true global citizen. I have been truly blessed to receive such mentorship. I will forever be grateful and never forget their investment.

Geoffrey Cohen, my undergraduate and Master's thesis advisor, who initially motivated me to pursue a Ph.D. in the United States; for his commitment and conviction that I should follow my inclinations to become a researcher and join a Ph.D. program in the United States.

Irene Blair, additional committee member and collaborator; for disagreeing with me, for respecting and at times even enjoying *my* disagreement with her, for her insight and her sense of humor, and for the countless interesting discussions, many of which are hopefully yet to come.

A variety of very talented and diligent research assistants have helped complete my doctoral work over the past five years. The following research assistants contributed especially to the work on this dissertation as lead experimenters of the studies presented: Jack Hager, Nicole Hein, and Brian Krantz for Study 1; A. Kismet Smith and Laura Durkin for Study 2; and finally, Owen Alexander and Daniel Milman for Studies 3, 4, and 5. The work on this dissertation would not have been possible without their hard work and dedication.

Part of my doctoral studies, and thus a significant amount of work on this dissertation, was supported by an ERP fellowship provided by the German National Academic Foundation and the German Ministry of Economy and Technology during the academic years 2009-2010 and 2010-2011. It was also supported by a summer dissertation finishing fellowship by the Graduate School of the University of Colorado in 2012. I am very grateful for the financial support I have received.

CONTENTS

TABLES

FIGURES

Introduction and Theoretical Background:

Introspection of Implicit Attitudes

Considerable interest in the concept of implicit attitudes has been shown over the past two decades in both academic outlets such as journals (e.g., Bosson, Swann & Pennebaker, 2000; Cunningham, Nazlek & Banaji, 2004; Jost, Pelham & Carvallo, 2002; Phelps et al., 2000; Rudman, Greenwald, Mellott & Schwartz, 1999; Quillian, 2008) or text books (Kassin, Fein, & Markus, 2008; Baumeister & Bushman, 2008; Kenrick, Neuberg & Cialdini, 2010); as well as more popular outlets such as popular books (Gladwell, 2005), newspapers (Tierney, 2008a, 2008b; The Economist, 2012), or even TV shows (Dateline NBC, 2007; Oprah.com, 2006). The attraction of implicit attitudes stems in large part from demonstrations that they capture aspects of human thought and behavior that are sometimes not revealed by self-report. For example, White respondents generally report their ethnic/racial attitudes as equally positive toward most groups, yet measures of their implicit attitudes reveal that they have substantial implicit preferences for Whites over other groups (e.g., Nosek, 2005, 2007; Divine, 1989). And importantly, these preferences can sometimes predict actual intergroup behavior better than people's self-reported (explicit) attitudes (e.g., Dovidio, Kawakami, & Gaertner, 2002; Fazio, Jackson, Dunton, & Williams, 1995; Rydell & McConnel, 2006; see Greenwald, Poehlman, Uhlmann, & Banaji, 2009).

Included in the discussion surrounding implicit attitudes has been a general assumption that people lack introspective access to these attitudes (Gawronski, Hofmann, & Wilbur, 2006). We highlight the circumstantial nature of the evidence supporting this assumption and review the theoretical claims made on the basis of it. Contrary to this assumption, and in line with prominent theoretical considerations (Gawronski & Bodenhausen, 2006; Gawronski et al., 2006),

we then present five studies in which we show that people can in fact introspect about their implicit attitudes with reasonable accuracy, even as they report distinct explicit attitudes.

The Assumption: Implicit = Unaware

The assumption that implicit attitudes are outside of conscious awareness can be found in both popular presentations and the academic literature. For example, the Project Implicit website (https://implicit.harvard.edu) – widely accessibly in more than 20 languages and averaging around 15,000 weekly visits –states, "[this] web site presents a method that demonstrates the *conscious-unconscious* divergences much more convincingly than has been possible with previous methods" (emphasis added; Nosek, Banaji, & Greenwald, 2006, last seen May, 2012, at the time of writing this manuscript). Widely used textbooks also describe implicit attitudes as unconscious attitudes that people "cannot self-report…because we are not aware of having them" (p. 207, Kassin et al., 2008; see also Baumeister & Bushman, 2008; Kenrick, et al., 2010). In the academic literature, the assumption that implicit attitudes are unconscious can be found in numerous articles, most often in the use of "implicit" and "unconscious" as interchangeable terms (e.g., Bosson et al., 2000; Cunningham et al., 2004; Jost et al., 2002; Phelps et al., 2000; Rudman et al., 1999; Quillian, 2008) but sometimes in stronger statements that implicit attitudes *cannot* be introspected (e.g., Devos, 2008; Greenwald & Banaji, 1995; Kihlstrom, 2004; Spalding & Hardin, 1999; McConnell, Dunn, Austin, & Rawn, 2011).

To be sure, there are various aspects of implicit attitudes that might be unknown to the individual possessor (Bargh, 1994; Gawronski et al., 2006). Gawronski and colleagues list three: the attitude's origin, the attitude's contents, or the attitude's impact on judgment and behavior. Our focus in this paper is on people's awareness of the contents of their implicit attitudes – in

part because the other two types of awareness (the attitudes' source and impact) would be impossible if people did not know what their implicit attitudes are (Gawronski et al., 2006).

Why are the contents of implicit attitudes assumed to exist outside of awareness or, in the strong form of this argument, why should it be impossible to know one's implicit attitudes? There are likely many answers to this question. We offer three that seem to be given with some regularity.

First, many people seem to be surprised about their results when they complete implicit attitude measures (e.g., Banaji, 2001; Gladwell, 2005; Nosek, 2007; Tierney, 2008b). People often attempt to "explain away" their results, presumably because the newly discovered implicit attitudes aren't consistent with what is consciously endorsed (Uhlmann & Nosek, in press; Uhlmann, Poehlman, & Nosek, in press). In other words, there is evidence that people often do not seem to think about their implicit attitudes before completing implicit-attitude measures, and this has been taken as evidence that implicit attitudes might not be available to introspection even if a person tried.

Second, other types of unawareness might imply a lack of content awareness, particularly a lack of awareness of the impact that implicit attitudes have on behavior. For instance, implicit attitudes have been shown to predict behavior during an interracial interaction as rated by outside observers, even though individuals often report different conscious perceptions of that behavior (Dovidio, et al. 2002; Fazio et al. 1995).

Third, and most prominently, there are by now a large number of studies that show implicit attitudes and explicit (inarguably conscious) attitudes are often only weakly correlated (Blair, 2001; Nosek, 2005, 2007; Nosek & Hanson, 2008; Hofmann, Gawronski, Gschwendner, Le, & Schmitt, 2005a; Hoffmann, Geschwendner, Nosek, & Schmitt, 2005b). If people's

introspections about their attitudes produce an answer (explicit attitude) that is different from what is revealed by implicit attitude measures, it seems reasonable to assume that the implicit attitude was not available for that introspective process. However, as we will try to show, according to theoretical models on implicit as opposed to explicit attitudes (e.g., Gawronski & Bodenhausen, 2006), explicit attitudes are not the pure result of introspection of implicit attitudes, and are influenced by a variety of other factors. For instance, many authors have noted the importance of self-presentational concerns for explicit reports (e.g., Hofmann et al., 2005a, 2005b; Nosek, 2005, 2007; Nosek & Hanson, 2008; Nosek & Smyth, 2007; Fazio et al., 1995). A large-scale study of 57 object pairs and a sample of 12,563 participants (Nosek, 2005) showed that self-presentational concerns do indeed moderate implicit-explicit attitude correlations. On the other hand, a meta-analysis by Hofmann et al. (2005a) confirmed that the impact of self-presentation is not particularly large and the field's emphasis on this factor as an explanation for discrepancies in implicit-explicit relationships may be exaggerated (see also Hofmann et al., 2005b).

In sum, there are several reasons why one may assume that implicit attitudes are not available to introspection: People are often surprised by their own results on implicit attitude measures; people appear to be often unaware of behavior that is based on implicit attitudes; and people's reports of their explicit attitudes are different from the attitudes that are revealed by implicit measures, and this difference cannot be fully attributed to self-presentational concerns.

On the other hand, the evidence supporting this assumption is circumstantial. No study has directly assessed people's ability to introspect on their implicit attitudes and thus direct disconfirmation is lacking. Further, the fact that implicit attitudes are unrelated to conscious

output (e.g., explicit attitudes) can be explained without making the further conclusion that people cannot access the contents of their implicit attitudes.

What Does Theory Say?

A consideration of the many theoretical models of implicit attitudes reveals a range of positions, from the statement that implicit processes are, by definition, introspectively unavailable (Greenwald & Banaji, 1995) to the claim that implicit attitudes are probably open to introspection, at least under certain circumstances (Gawronski & Bodenhausen, 2006; Wilson, Linsey, & Schooler, 2000) – with most models simply not addressing this issue (Bodenhausen & Macrae, 1998; Devine, 1989; Smith & DeCoster, 2000; Strack & Deutsch 2004). We focus here on Gawronski and Bodenhausen's (2006) Associative-Propositional Evaluation (APE) model, which motivated the current research.

According to the APE model, implicit attitudes are based on spontaneous affective reactions that occur in response to an attitudinal cue and, importantly, have no "truth value" assigned to them (Gawronski & Bodenhausen, 2006). That is, these associations are activated in response to a stimulus (or an array of stimuli that form a category) regardless of whether the perceiver believes that the associations are valid or invalid. For example, many White Americans appear to have spontaneous negative associations with Black Americans, even when they regard that negativity as invalid or false (Divine, 1989).

Now, when the same people who might have spontaneous negative associations with Black Americans are asked how they "feel" about this group (i.e., asked to indicate their explicit attitude), a whole different process is set in motion according to the APE model. Specifically, explicit attitudes result from an inferential process that considers all of the propositions or statements that are activated at the time the judgment is made, and that are considered relevant

for an evaluation of the attitude object. These propositions may reflect specific exemplars of a category that come to mind ("I really like my friend Martin who is Black;" "I really like Bill Cosby."), but may also include other sources, such as one's values (e.g., "I believe that all people are fundamentally good." "I should not like or dislike entire groups of people based on race or ethnicity"); other relevant knowledge that can cause affective reactions ("Disadvantaged groups really have a rough time in society." "I admire the fight certain groups have fought for their rights"); or, as mentioned, self-presentational concerns ("I shouldn't say that I have negative feelings towards social groups even if I feel that way"). One may also consider propositions based on spontaneous associations in propositional form ("I initially feel uncomfortable when I think about Blacks.").

The most important aspect of this inferential process, for present purposes, is a determination of a propositions' truth-value: Which thoughts and feelings are considered valid for judgment and which ones are considered invalid for this purpose? Specifically, a person will try to create consistency between the different propositions (Gawronski & Bodenhausen, 2006; Gawronski, Brochu, Sritharan, & Strack, 2012): How can the different propositions, some of them causing negative reactions and some positive, be integrated to form a consistent explicit affective judgment?

As implied in the foregoing example, a determination that one's spontaneous associations are inconsistent with other propositions and thus invalid will result in their exclusion from the explicit attitude. In this case implicit and explicit attitudes will differ. On the other hand, when a person considers his or her spontaneous associations (implicit attitudes) to be a valid basis for judgment, then these associations will be considered as part of the formation of an explicit attitude. However, even in that case, other propositions that are also seen as valid will be

considered as well. In sum, the explicit attitude will be a result of this process of bringing different propositions to mind, weighing them against each other, and creating consistency among them, rather than a pure result of introspection of an implicit attitude.

The most important point we wish to make regarding the APE model and its application is that lack of awareness of one's implicit attitude is not a necessary condition for implicit and explicit attitudes to mis-align. Indeed, one may be well aware of one's spontaneous negative associations and then consciously decide against them, as suggested in the prior example. In other words, low correlations between implicit and explicit attitudes say little about a person's awareness of their implicit attitudes. To the contrary, it appears that the only way to determine whether or not people are able to introspect their implicit attitudes is to directly ask them about these implicit attitudes.

Although previous research has not directly asked people to report their implicit attitudes, research on the relationship between implicit and explicit attitudes is consistent with the possibility that people may be able to do so. Specifically, numerous studies have shown that explicit attitudes are more strongly related to implicit attitudes when participants are asked to listen to their intuitive gut reactions and make fast judgments as opposed to taking their time and making deliberate judgments (e.g., Smith & Nosek, 2011; Ranganath, Smith, & Nosek, 2008). For instance, In a study by Ranganath et al. (2008), participants were presented with attitude objects and asked to report their gut feelings and their actual feelings, their instant reactions and fully considered attitudes, and to complete measures of implicit attitudes. Ranganath et al. (2008) found that reports of gut feelings and instant reactions loaded more strongly with implicit attitudes than with reports of actual and fully considered feelings (i.e., explicit attitudes). These studies demonstrate that people can report gut/instant reactions that reflect on their implicit

attitudes to some degree. On the other hand, none of these studies directly asked participants to report an implicit attitude or predict a score on an implicit attitude measure, and the higher correspondence could thus have different reasons. Hence, direct confirmation that implicit attitudes can be introspected upon is still lacking.

The Present Research

Our research introduces a paradigm in which participants are asked to predict their own results on measures of implicit attitudes toward five different social groups. After participants complete the tests, we examine the degree to which their predictions correspond to their actual results across the five IATs, as well as with their self-reported explicit attitudes. Based on considerations in line with the APE model (Gawronski & Bodenhausen, 2006) and other suggestive findings (e.g., Ranganath et al.'s, 2008), we hypothesized that participants would be reasonably accurate in predicting their implicit attitudes. We also predicted that participants would likely report explicit attitudes at the same time that are distinct from both their IAT predictions and their actual IAT scores.

We asked participants to complete five different IATs (and make predictions for these) for two different reasons. First, we wanted to have a range of valenced target objects where different propositional considerations might be relevant. So, for instance, we included IATs that captured traditional implicit prejudice (e.g., White versus Black targets) and IATs where normative concerns were likely to be irrelevant (e.g., children versus adult targets).

The second reason for our use of five different IATs was based on theoretical considerations about the appropriate unit of analysis in considering the accuracy of a person's prediction of their implicit attitudes. One way to analyze the relationship between implicit attitudes and implicit-attitude predictions would be to examine whether participants can predict

their implicit attitudes towards a particular attitude object relative to other participants. This would involve correlating predictions and implicit attitude scores across participants (between-subjects), one attitude object at a time. Another possibility would be to examine whether participants can predict whether their implicit attitudes are relatively high or low for one attitude object compared to other attitude objects. This latter method would mean correlating predictions and implicit attitude scores within a person across objects – effectively estimating a correlation for each participant separately across the five attitude targets (within-subjects). We believe that this second way (estimating within-subjects correlations) of examining the accuracy of implicit predictions is theoretically preferable.

Specifically, to make relatively accurate predictions of implicit attitudes when accuracy is assessed between-subjects for each object at a time (the first possibility described above), participants need to have access not only to their own implicit attitudes but also to knowledge of where their own implicit attitudes stand relative to others' implicit attitudes. While this latter question is interesting, it is altogether a different question. We were primarily interested in whether or not participants can introspect upon their own implicit associations, rather than their correctness in estimating how these attitudes would compare to the attitudes of others'. For these reasons, in the studies that follow, we examine the accuracy of introspection by correlating IAT score predictions and actual IAT scores within-subjects across the five IAT's that each participant completed. We then aggregate this relationship in a multi-level analysis to see whether participants' predictions are on average accurate, across participants. We return to the question of estimating how one's implicit attitudes compare to other people's implicit attitudes towards the end of this paper, in a separate analysis across studies.

Second, we pursued the additional question of self-presentational concerns altering reports. As discussed earlier, one of the explanations given for discrepancies between implicit and explicit attitudes is that people distort the latter when an honest attitude report would reflect poorly on them (e.g., negative attitudes toward minority groups in the face of egalitarian social norms). On the one hand we did not believe that participants would intentionally distort their IAT score predictions because we told them that their IAT scores would actually be measured – undermining the social advantages of an obvious lie. However, we reasoned that participants might not want to report certain attitude scores for themselves because of their mis-match with desired self-concepts (Gawronski et al., 2006, Wilson et al., 2000). To examine this possibility, half of the participants in Study 1 were informed that implicit attitudes are really cultural associations that may or may not reflect their true selves (removing self-concept threat), whereas the other participants were told that implicit attitudes are their true attitudes. If introspection on implicit attitudes is vulnerable to self-enhancing repression (Gawronski et al., 2006; Wilson et al., 2000), we reasoned, then participants ought to make worse predictions about their implicit attitudes in the "true attitudes" condition than in the less threatening "cultural associations" condition (Uhlmann & Nosek, in press).

Study 1

Method

Participants and Design. Sixty-nine undergraduate students participated in the study for partial course credit. One participant did not complete the measures and three more participants made too-fast responses (< 300 ms.) on more than 10% of their IAT trials and were thus excluded in accordance with criteria outlined by Greenwald, Nosek, and Banaji (2003). For the remaining 65 participants, 54% were women, and 82% self-identified as White. The other

ethnic/racial identities were, "other" or multi-racial (5), Latino (3), Asian (2), Black (1), and Middle-Eastern (1). Ages ranged from 18-25, with a median age of 19.

This study used a multi-level design. The continuous relationship between the five IAT score predictions and the five IATs were modeled at level one for each participant. The outcome of this relationship was modeled across participants at level 2. At level 2 we additionally assessed the influence of two differing IAT explanations on the strength of this continuous relationship.

Materials.

The IATs. Participants completed five evaluative IATs in random order, each comparing a different social group with the same reference group (White young-adults). The comparisons were labeled as, *Blacks* vs. *Whites*, *Latinos* vs. *Whites*, *Asians* vs. *Whites*, *Celebrities* vs. *Regular People*, and *Children* vs. *Adults*. All pictures representing children and celebrities were also White to ensure the social dimension was perceived as intended.

Ten faces (five male and five female) representing each social group were selected from the productive ageing lab database (Minear & Park, 2004) and from photos found publicly available online. Each face had a neutral expression, included the person's hair and neck, and was shown against a grey background. The pictures used in each IAT were pretested and matched on likeability, except for those in the categories "child" and "celebrity" which were not expected to be comparable in liking to average White adults. The faces used to represent the reference group (White young adults) were different in each IAT, thus there was a total of 50 White young adult faces used.

Each of the five social group IATs consisted of the following four blocks: (1) 20 trials sorting pictures of the two social groups, (2) 40 trials in which one group was sorted with

positive words and the other group was sorted with negative words, (3) 40 trials in which the two social groups were reversed in position from Block 1, and (4) 40 trials in which the groups were paired with the opposite valence from Block 2.[1] To ensure comparability between participants, all participants always received compatible blocks first, and incompatible blocks second (i.e., the order was not counterbalanced, Hofmann, Gschwendner, Wiers, Friese, & Schmitt, 2008; Egloff & Schmukle, 2002; Gawronski, 2002). For all of the IATs, the *compatible* blocks were defined as those blocks in which the reference group is paired with good words, i.e., White+good for the three ethnic/racial IAT, Adult+good for the Children:Adult IAT, and RegularPeople+good for the Celebrities:Regular People IAT.

An IAT *D*-score following recommendations by Greenwald et al. (2003) was calculated for each person on each IAT (i.e., the difference between the incompatible and the compatible blocks divided by their pooled standard deviation for each participant). Higher scores on this measure reflect less positive implicit attitudes toward each social group (i.e., Blacks, Latinos, Asians, children, or celebrities) compared with the reference group.

IAT as True Attitudes vs. Cultural Associations. Several steps were taken to manipulate beliefs that the IAT reveals either true attitudes or cultural associations. In the true attitudes condition, the participants were first given a half-page introduction in which the IAT was described as revealing a person's "true underlying attitude" that can sometimes differ from what people "think of themselves." After they had read this information, participants were asked to explain in their own words the difference between implicit attitudes and explicit attitudes.

---

[1] Participants also completed two shorter practice IATs described below. The valence words were sorted alone during the first practice IAT. Because good and bad words were sorted the same way for all following IATs, this block was not repeated in any of the social-group IATs afterwards.

[2] All of the multi-level analyses were conducted using the mixed-model commands in SPSS/PASW 19 with its associated default settings for statistical conventions, unless otherwise noted.

[3] In the model with standardized values, each participant's mean IAT score, and thus the level-one intercepts equal zero. Accordingly, these intercepts do not vary and cannot be modeled as function of condition.

[4] As can be seen, the size of the relationship dropped dramatically in this analysis. We believe that the re-

Participants were next given more specific information about the IAT procedures that likened the tasks to sorting cards. All references to IAT results in this section were consistently labeled as "true implicit attitudes," and again the participants were asked to write down what they had just learned, using their own words.

Conversely, participants in the cultural associations condition received a description of the IAT as revealing "culturally learned associations" that can differ from "what the person truly believes," and all references to IAT results were phrased as "culturally-learned associations." These participants were also asked to write down what they had learned about implicit associations and the IAT, following each description.

After learning that the IAT reveals "true implicit attitudes" or "cultural associations," all of the participants were given first-hand knowledge of the IAT by completing two practice tests; one comparing Insects vs. Flowers and the second comparing Dogs vs. Cats. These targets were chosen because we thought that the IAT results they produced would be believable as either true attitudes (likely consistent with their explicit feelings toward the targets) or as cultural associations (most people like flowers over insects, and there are enough dog and cat lovers to make either attitude seem normative as well). The practice IATs had only half the number of trials of the regular IATs to give participants a good sense of the test but not fatigue them unnecessarily. For both of the practice tests the participants were asked to first predict their score, complete the IAT, indicate again how they thought they had scored, and then received automatized feedback on their actual IAT score.

IAT prediction task. Prediction of one's own performance on an IAT was asked in terms of the perceived "ease" of completing the compatible versus the incompatible sorting tasks. For example, in predicting their performance on the Black-White IAT, participants were shown the

faces that would appear in this test with one group appearing above the left side of the 7-point response scale and the other group appearing above the right side of the scale. Participants were encouraged to look at the pictures, "carefully listen to their gut feeling," and then try to answer the question of which sorting task (sorting Black with good or sorting White with good) would be easier for them, and how much easier it would be (see *Figure 1*). This bi-polar scale thus made the comparative nature of the IAT clear and asked the participants to respond accordingly.

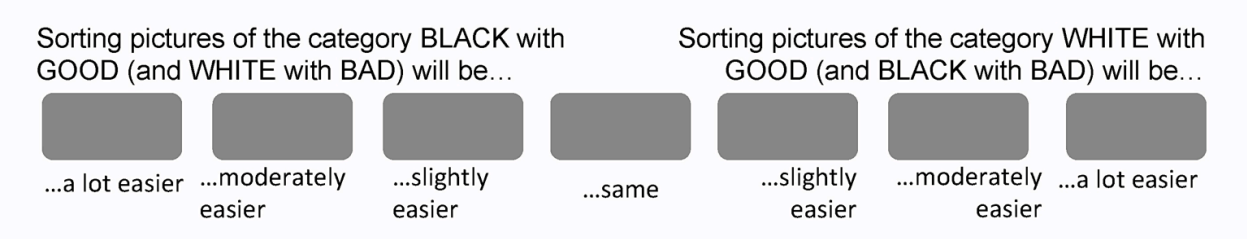

Figure 1: Prediction scale participants used to make their predictions of their IAT score (example of Black-White IAT) in Study 1. Photos used in the actual IATs were depicted above the ends of the scales on the left and right. In Study 2, labels below the buttons were changed (see text).

When participants received feedback on their performance, the computed *D*-scores were translated into terms that were similar to the prediction scale: *D*-scores >.65 produced the feedback that a particular sorting combination had been "A LOT easier" than the other combination, *D*-scores between .65 to .35 were translated as "MODERATELY easier," D-scores between .35 and .15 were translated as "SLIGHTLY easier," and D-scores between .15 and -.15 produced the statement that the two sorting tasks were "the SAME" for the participant. These cut-offs were made according to conventions used on the IAT webpage (Personal communication from N. Sriram to I. Blair on July 6, 2009).

Explicit Ratings. Participants were asked to indicate their explicit group attitudes using a standard thermometer scale. For each group label, a scale appeared on the computer screen in

the shape of a thermometer and ranged from "0 – very coolly" to "100 – very warmly." Participants were asked to indicate how warmly or coolly they felt toward "Whites/Caucasians," "Blacks/ African Americans," "Asians/ Asian Americans," "Latinos/ Hispanic Americans," "children," and "celebrities."

Manipulation Check. To assess the extent to which participants in each condition had in fact accepted the explanation they were given about the IAT, they were presented with four statements about the IAT, each accompanied by a seven-point scale ranging from "1 – strongly disagree" to "7 – strongly agree." The statements were, "The IAT measures my true underlying attitude", "My IAT results have nothing to do with how I really feel about different groups of people" (reverse-scored), "The IAT measures a culturally learned association that I hold," and "The IAT cannot say how I'm influenced by my culture" (reverse-scored). The first two and last two items were averaged, respectively (*Cronbach's α* attitudes scale = .62, *Cronbach's α* associations scale = .47), with higher scores indicating more agreement with the respective explanation.

Procedure. After informed consent was obtained, participants were seated in individual cubicles and completed the tasks in the following order: (1) explicit thermometer ratings, (2) explanation of the IAT (true attitudes vs. cultural associations, randomly assigned), (3) two practice IATs, each with a prediction, a post-diction, and computer feedback about the actual result, (4) predictions of their IAT scores for *all five* of the critical IATs in one pre-determined order (Black-White, Asian-White, Latino-White, children-adults, celebrities-regular people), and (5) completion of the five IATs in random order. After completing each IAT, participants were asked to re-assess ("post-dict") their score. The experiment concluded with participants making

explicit thermometer ratings again on all of the groups. They then answered the manipulation check and demographic questions.

Results

Manipulation check. Before the primary analysis, we examined the manipulation check responses to determine whether participants had accepted our manipulation of IAT explanation. We ran a 2 (condition: true implicit attitudes vs. cultural association condition) by 2 (scale: IAT measures culturally-learned associations vs. IAT measures true implicit attitudes) mixed-model ANOVA with repeated measures on the second factor. The expected interaction of the two factors emerged, $F(1, 63) = 31.08$, $p < .00001$. Participants in the true-attitudes condition (M=5.08, SE=.20) agreed significantly more than participants in the cultural-associations condition (M=4.15, SE=.21) with the idea that the IAT measured their true underlying attitude, $F(1, 63) = 10.08$, $p = .002$. Conversely, participants in the cultural associations condition agreed significantly more with the idea that the IAT measured culturally learned associations (M=5.06, SE=.19) than participants in the true-attitudes condition (M=4.40, SE=.18), $F(1, 63) = 6.36$, $p = .014$. Hence, participants believed that the IAT measured what we told them it measured – either their true underlying attitudes, or culturally learned associations that they hold.

IAT scores. *Figure 2* depicts the mean IAT D scores. As expected, on average participants tended to have more positive implicit attitudes towards Whites as compared to Blacks, Latinos, or Asians, all three $t(64)$'s > 7.5, all p's < .00001, but more positive attitudes towards celebrities as opposed to regular people, $t(64) = -2.35$, $p = .02$, and children as opposed to adults, $t(64) = -3.74$, $p < .001$.

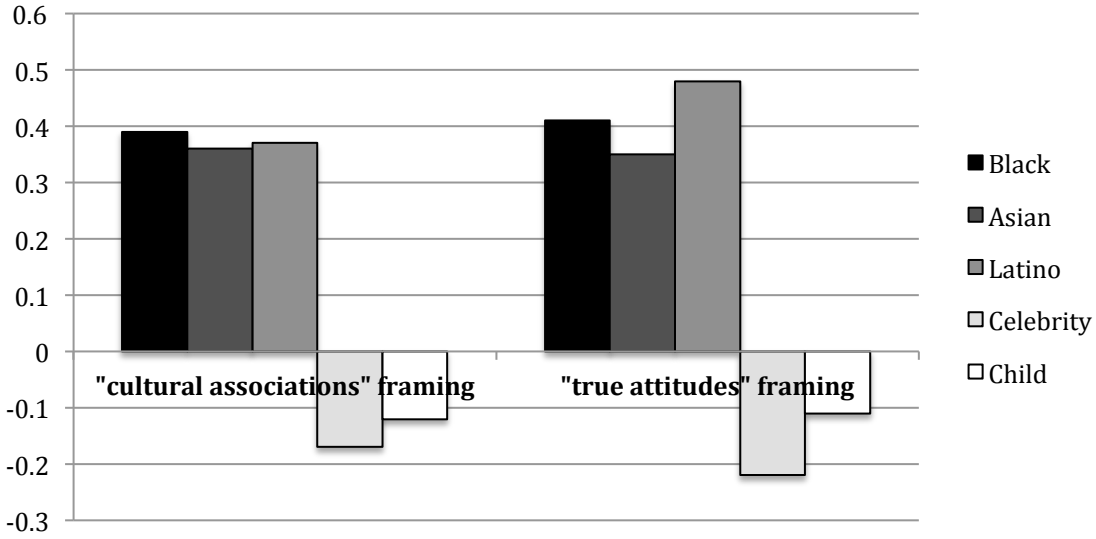

Figure 2: Study 1: Mean IAT scores by condition. Higher scores mean more positive implicit attitudes towards the comparison group (i.e., Whites, adults, or regular people). Negative scores indicated more positive scores towards the target group (Blacks, Asians, Latinos, Celebrities, or Children).

Accuracy of predictions. In order to examine whether participants could accurately predict their five IAT scores, we estimated a multilevel model[2] in which each participant's five IAT scores were modeled as a function of the person's five IAT predictions at the first level. The resulting slopes of this analysis estimate the degree to which IAT scores are associated with participants' predictions. Because we wanted the sizes of these random slopes to be representative of participants' accuracy in predicting the patterns of their results, we individually person-standardized each IAT score and each prediction by participant. The slopes are thus akin to a correlation coefficient for each participant. At level two, between-participants, we looked at the average size of these random slopes (the fixed effect), and also modeled them as a function of the IAT explanation condition. The results from this analysis are given in the left column of

[2] All of the multi-level analyses were conducted using the mixed-model commands in SPSS/PASW 19 with its associated default settings for statistical conventions, unless otherwise noted.

Table 1.[3] The top part of Table 1 gives the tests of the fixed effects and the bottom shows the variances of the random error components of the model.

Table 1: Study 1: The effect of IAT score prediction and explicit thermometer ratings on actual IAT scores by condition

Parameters	Score predictions	Explicit therm. ratings
Fixed effects		
IAT score predictions	.53***	
Predictions × condition	.09†	
Explicit therm. ratings		.01
Random effect variances		
IAT score predictions	.041	
Explicit therm. ratings		.077
Residuals	.550***	.741***
Goodness of fit		
-2 log likelihood	750.43	850.39

†p < .11 **p<.01 ***p<.001

The dependent variables in both models are participants' IAT scores. Both level-1 variables, the dependent IAT scores and the predictions, are standardized for each individual participant before they are entered in the analysis. Accordingly the intercept in this model would be 0 and is not included in the model. Similarly, the main effect of condition on these centered IAT scores is not included either.

As Table 1 shows, participants' predictions of their IAT scores corresponded significantly with their actual IAT scores, b= .53, $t(61)$ =9.80, p <.001. As said, the slope from this standardized model can be interpreted as the average within-participant correlation between predictions and actual IAT scores. Looking at the distribution of these correlations separately revealed that it was negatively skewed, making the median within-participant correlation between predictions and IAT scores somewhat higher, r = .62.

As Table 1 also indicates, the manipulation of IAT explanation did not affect mean IAT scores and only minimally affected the prediction slopes (Predictions by Condition interaction).

[3] In the model with standardized values, each participant's mean IAT score, and thus the level-one intercepts equal zero. Accordingly, these intercepts do not vary and cannot be modeled as function of condition.

Contrary to a threat hypothesis, score predictions were actually somewhat more accurate in the "true-attitudes" condition than in the "associations" condition, b= .09, $t(60)$ =1.64, p =.11.

The lack of condition effect made us wonder whether or not the manipulation possibly only affected the IATs run on minority targets. That is, a revelation of "true attitudes" should be most unpleasant when it reveals possibly un-egalitarian views on minorities, whereas it might be irrelevant in a prediction of attitudes towards children or celebrities. Results did not support this explanation. First, across conditions there was still evidence that participants could predict their pattern of these three scores, even in this underpowered analysis based on only three data points per participant, b=.32, $t(38.0) = 2.80$, p=.008.[4] More importantly, there was still no evidence that thinking about the IAT as revealing "true attitudes" made participants less accurate. As before, the direction of the slope, if anything, indicated more accuracy in the "true attitudes" as opposed to the "cultural associations" condition, b=.15, $t(38) = 1.29$, p=.21.

Relations with explicit attitudes. An additional analysis was conducted to determine whether participants' explicit attitudes could also predict their IAT scores. This time our level-1 modeled IAT scores as a function of participants' explicit attitudes (their thermometer ratings) measured prior to the actual IATs.[5] The results of this analysis are reported in the right column of Table 1. They show that participants' explicit attitudes were unrelated to their implicit attitudes across the five IATs, b = .01, $t(64)$ =.10, p =.92.

---

[4] As can be seen, the size of the relationship dropped dramatically in this analysis. We believe that the reduction in size of the correlation can be mostly attributed to lack of statistical power. For instance, 25 of the 65 participants made the same predictions for all three minority IATs on our restrictive 7-point prediction scale, and their correlation values were thus inevitably 0, even though these could have theoretically been accurate predictions of their pattern of results within the constraints of the prediction scale. The size of the interaction slope (.15) furthermore indicates that the mean correlation was .47 in the attitudes condition, but only .17 in the associations condition on average, even though this difference was not significant, further questioning the validity of interpreting these correlations based on only three data points meaningfully.

[5] The model does not include Condition as a level-2 predictor because the participants were randomly assigned to condition *after* they gave their explicit thermometer ratings.

To examine whether the experience of taking an IAT may alter a person's self-reported explicit attitude – either because it is viewed as truly informative or participants realize that we have "objective" evidence about their attitudes – participants' explicit attitudes reported at the end of the study were modeled as a function of their IAT scores. The results of this analysis are given in the left column of Table 2, labeled M1. Unlike the pre-IAT explicit attitude measures, the post-IAT explicit attitude measures were in fact related to IAT scores, b = -.28, $t(63)$ =-4.07. p <.001.[6] This relationship did not depend on explanation condition, both t's < .5. That is, participants did show adaptation of their explicit attitudes to their implicit attitudes after the introspection procedure.

Table 2: Study 1: Explicit thermometer ratings assessed after participants completed all IATs, regressed onto participants' IAT D scores.

Parameters	M1 – simple model	M2 – controlling for predictions
Fixed effects		
IAT *D* scores	–.28***	.01
IAT *D* scores × cond.	–.02	.05
IAT score predictions		–.53***
Predictions ×condition		–.04
Random effect variances		
IAT *D* scores	.144**	.134**
IAT score predictions		.027
Residuals	.619***	.463***
Goodness of fit		
-2 log likelihood	814.17	743.26

†p < .07 *p<.05, **p<.01 ***p<.001

The dependent variables in both models are participants' explicit thermometer ratings assessed after the IAT procedure. All level-1 variables, including the dependent IAT *D* score, are standardized for each individual participant before they are entered in the analysis.

Interestingly, once controlling for participants' predictions of their IAT scores, the relationship between (post-IAT) explicit thermometer ratings and IAT scores disappeared, b =

[6] Note that the negative sign is in line with predictions since on a thermometer higher values mean more positive evaluations, but on the IATs higher scores represent more pro-White bias.

.01, $t(82.35) = .12$, $p = .90$ (see the right column of Table 2). That is, whereas the actual experience of completing the IAT had a significant influence on participants' self-reported attitudes at the end of the session, this influence could be entirely explained through their awareness (predictions) of what their implicit attitudes would be.

Discussion

The purpose of Study 1 was to investigate people's ability to introspect on the content of their implicit attitudes, measured here with the IAT. With regard to five social-group comparisons, we found that participants did indeed predict their implicit attitudes with a fair amount of accuracy (median $r = .62$). Although participants had practice with the IAT procedures before making their predictions, they were not given any IAT experience with regard to the specific social-group attitudes they were asked to predict.

Of further interest was the finding that participants were as accurate (even a little more so) in their predictions when they were told that the IAT revealed their true attitudes, as they were when they were told that the IAT instead revealed culturally learned associations. Results from the manipulation check showed that participants accepted the explanations they were given. Taken together, the pattern of results suggests that people can introspect their implicit attitudes – rather than shutting-down or repressing those feelings – even as they face the possibly unpleasant revelation that their "true attitudes" differ from their conscious beliefs.

We believe the study findings are most consistent with the premises of the APE model of attitudes (Gawronski & Bodenhausen, 2006): (a) Implicit attitudes represent spontaneous evaluative associations of which the perceiver can be aware; (b) awareness of these associations does not mean that they will be incorporated into the perceivers' explicit attitude, which may be

based on a variety of propositions, and (c) conversely, a lack of correspondence between implicit and explicit attitudes does not imply that people are not aware of their implicit attitudes.

Not only did participants predict their implicit attitudes with reasonable accuracy, but they did so at the same time that they reported explicit attitudes that were *not* related to their implicit attitudes. Once participants were given the opportunity to revise their explicit attitudes in light of their IAT performances, the increased correspondence between these post-IAT explicit attitudes and implicit IAT scores was entirely explained by the participants' *a priori* predictions about their IATs (i.e., the insights they had on their implicit attitudes even before they learned their IAT results).[7]

Despite the consistency of the results with the APE model, there are additional questions that remain to be answered. Specifically, the prediction task in Study 1 asked participants to make a rather operational prediction (i.e., "Which of two blocks in this reaction time task will be easier to complete?" See *Figure 1*). Participants might have some sort of procedural awareness about their response impulses, but not about their spontaneous affective reactions towards the groups. Since it is the latter construct (feelings) that we intended to address, Study 2 was conducted to replicate the results with a more conceptual prediction measure about their spontaneous feelings towards the groups rather than their response impulses on the IAT.

Study 2

Method

Participants and Design. Ninety-three undergraduate students participated in the study for partial course credit. Three participants were excluded from data analysis: One participant made too-fast responses (<300 ms) on 19% of the IAT trials (Greenwald et al., 2003), and two

[7] We found the same results when we conducted these analyses only with the three ethnic-group IATs. That is, despite the reduction in power of calculating correlations with only 3 data points per participant, the pattern of results was replicated with only the ethnic-group IATs in this study.

participants were missing too much data to be included. The remaining 90 participants were 64% women, and 81% identified as White. The other ethnicities were: 5 "other" or mixed-races, 5 Arab/Middle-Eastern, 4 Asian, 2 Latino and 1 Black. Ages again ranged from 18-25, with a median age of 19.

Study 2 used the same design as Study 1. That is, the continuous relationship between IAT scores and IAT score predictions was calculated for each participant at level 1 (based on values standardized individually for each participant), and the effect of the IAT explanation condition on this relationship was assessed across participants at level 2.

Materials and procedure. The materials and the procedure were exactly the same as those used in Study 1 with two exceptions. First and most significantly, we modified the measure that the participants used to predict their IAT results. As before, participants saw the pictures they would be sorting in the IATs accompanied by instructions encouraging them to look at the pictures and listen to their gut reactions. However, instead of focusing on which of two IAT blocks would be easier to complete, the prediction measure asked participants directly about their "true implicit attitudes" or their "culturally learned associations", depending on IAT-explanation condition. Thus, the final prediction scale read, e.g., "I predict that the IAT comparing my reactions to BLACK vs. WHITE will show that my true implicit attitude is…" (1) "a lot more positive towards BLACK", (2) "moderately more positive towards BLACK", (3) "slightly more positive towards BLACK", (4) "same", and then the opposite labels on the second half of the scale (e.g., "slightly more positive towards WHITE," etc.). Except for these changes in the labels, the prediction scale still looked similar to the one depicted in *Figure 1*.

The second modification was made to try and reinforce and strengthen the assigned explanation of the IAT as revealing true implicit attitudes or culturally learned associations,

given its weak effect in Study 1. Following score feedback on the two practice IATs (insects-flowers and dogs-cats), participants were asked to write again. This time they were asked to reflect about the two practice IATs and the scores they had shown. In line with condition assignments, they were asked to write about what their scores had revealed about their "true attitudes" or their "culturally learned associations," respectively. As in Study 1, the effectiveness of the manipulations was assessed in the end (*Cronbach's α* attitudes scale = .51, *Cronbach's α* associations scale = .48).

Results

Manipulation check. The manipulation check scales were analyzed as a function of IAT-explanation condition. The predicted interaction between scale and condition emerged again, $F(1, 88) = 11.14$, $p = .001$. Participants in the true attitudes condition agreed significantly more that the IAT measures their underlying true attitudes (M=4.54, SE=.21) than participants in the associations condition (M=3.90, SE=.18), $F(1, 88) = 4.87$, $p = .03$. Conversely, participants in the associations condition agreed marginally more that the IAT measures culturally learned associations (M=4.51, SE=.18) than participants in the attitudes condition (M=4.04, SE=.21), $F(1, 88) = 3.30$, $p = .07$.

Accuracy of predictions. As in Study 1, a multi-level model was estimated in which each participant's five IAT scores were modeled as a function of the person's five IAT predictions (individually person-standardized) at the first level, within-subjects, and then the random slopes[8] from this level were modeled at level 2 as a function of the IAT explanation condition, between-subjects. The results are summarized in the left column of Table 3. We again found a significant relationship between the IAT predictions and actual scores, $b = .55$,

[8] As in Study 1 and nearly all analyses presented in this paper, the within-participant standardization leads to all intercepts being 0. Hence, these intercepts do not vary randomly and cannot be modeled as a function of condition on level 2.

$t(86) = 12.02$, $p < .001$. The distribution of the correlation coefficients was even more negatively skewed in this study, and the median correlation was even higher than in Study 1, $r = .72$.

Table 3: Study 2: The effects of predictions and explicit thermometer ratings on IAT D scores.

Parameters	Prediction model	Explicit-implicit model	Simultaneous regression
Fixed effects			
IAT score predictions	.55***		.59***
Predictions × condition	–.00		–.04
Explicit therm. ratings		–.20***	.06
Therm. × condition			–.09
Random effect variances			
IAT score predictions	.048		.040
Explicit therm. ratings		.028	.079*
Residuals	.533***	.747***	.475***
Goodness of fit			
-2 log likelihood	1027.56	1161.51	1023.31

**$p<.01$ ***$p<.001$

The dependent variables in all models are participants' IAT *D* scores. All level-1 variables, the dependent IAT *D* scores and the predictions, are standardized for each individual participant before they are entered in the analysis.

An examination at level 2 of the model showed that despite our attempts to further reinforce the two different explanations for the IAT, there was no evidence that this manipulation had any effect on the accuracy of participants' predictions, seen in the slope value of the interaction of predictions by condition assignment, $b = .00$, $t(86) = -.09$, $p = .93$.[9]

Explicit attitude relations. The next multi-level model tested whether participants' explicit thermometer ratings were related to their IAT scores. Results are summarized in the

[9] As in Study 1, we also looked at whether the condition effect would show on only the minority IATs. This time, there was a marginal effect of condition*predictions in this analysis based on only three IATs per participant, $b=.17$, $t(45) = 1.75$, $p=.09$. However, as in Study 1, the direction of this slope was opposite to a threat hypothesis. That is, participants were, if anything, more accurate in predicting the pattern of only their minority IAT scores when they thought of them as revealing "true attitudes" than when they thought of them as revealing "cultural associations." The overall relationship between predictions and actual IAT scores was comparable to the one found in Study 1, $b=.33$, $t(45)=3.46$, $p=.001$. The number of participants giving the same prediction for all three IATs was 43 out of the 90 participants this time.

middle column of Table 3. This time the explicit thermometer ratings taken at the start of the study were significantly related to the IAT scores, b = -.20, t(89) =-4.12, p < .001. Importantly, however, this relationship was weaker than the relationship between IAT score predictions and IAT scores (.55 opposed to .20, see Table 3). Additionally, this relationship disappeared once participants' IAT predictions were added to the model as a covariate, b = .06, t(104.27) =1.09, p =27 (see right-most column in Table 3).

As in Study 1 we also wanted to see whether participants would adapt their explicit ratings after completing the five IATs. Results are noted in Table 4. As in Study 1, there was a significant relationship between IAT scores and the explicit attitude measures taken after the IATs that was larger than the relationship between IAT scores and explicit measures taken before, b = -.35, t(88) =-7.41, p <.001. However, note again, that this relationship was substantially lower than the relationship between predictions and IAT scores (b = -.35 on the latter as opposed to .55 on the former). And importantly, this relationship again disappeared when IAT predictions were added to the model as a covariate, b = -.05, t(118.52) =-1.15, p =25.

Table 4: Study 2: Explicit thermometer ratings assessed after participants completed all IATs, regressed on participants' IAT D scores.

	M1 – simple model	M2 – controlling for predictions
Fixed effects		
IAT *D* scores	–.35***	–.05
IAT *D* scores × condition	–.06	–.14**
IAT score predictions		–.53***
Predictions ×condition		.10†
Random effect variances		
IAT *D* scores	.040**	.026
IAT score predictions		.100**
Residuals	.644***	.432***
Goodness of fit		
-2 log likelihood	1105.37	985.22

†p = .09 **p<.01 ***p<.001

The dependent variables in both models are participants' explicit thermometer ratings assessed after the IAT procedure. All level-1 variables, including the dependent IAT *D* score, are standardized for each individual participant before they are entered in the analysis.

In other words, participants' explicit attitudes in this study appeared to be informed, in part, by their spontaneous associations (implicit attitudes), and more so after participants completed the IATs. However, this relationship could be explained entirely by participants' awareness of these associations.

Discussion

Study 2 replicated the results of Study 1 with a more conceptual predictions task. Instead of predicting which of two blocks would be easier to complete in an IAT, Study 2 asked participants to predict their "implicit attitudes". Participants were equally accurate (even slightly more accurate) at making this prediction as they were in Study 1. The relationship between implicit and explicit measures again followed a different pattern. While we observed a significant relationship between implicit and explicit attitudes in Study 2, this relationship was a) substantially weaker than the relationship between implicit attitudes and their predictions, and b) could be entirely explained through participants' introspections (predictions) of their implicit attitudes. We believe this to be consistent with the APE model: Participants notice their implicit attitude and consider it as one of many propositions that will factor into their explicit attitude. Because different participants will make different decisions, and because even those participants who consider their implicit attitude a valid contributor to an explicit judgment will align it with other propositions, the average relationship will remain low.

Results so far confirmed that participants can accurately predict their implicit-attitude scores. But how do we know that these accurate predictions are in fact a result of an introspection process? Looking at the groups that participants are asked to evaluate in our study,

another possible explanation is that participants simply predict the IAT pattern that "makes the most sense." For instance, most contemporary Americans would probably predict that the average American would have somewhat more negative associations with ethnic minorities as opposed to Whites, but somewhat more positive associations with children as opposed to adults and celebrities as opposed to regular people. The presumed "accuracy" in our predictions could be proof to this simple fact that participants predict "reasonable" responses, based on their naïve theories about social norms, and then largely confirm to those, rather than an account of true introspection (F. Strack, personal communication, July 2011). We decided to start a first attempt at investigating the difference between true introspection and such theories about social norms. Specifically, we reasoned that participants' ideas about normative responses should be more or less the same across participants. Their average prediction should thus represent more or less what participants on average believe the pattern of responses should be.

We thus ran one additional analysis on the data of Study 2 comparing participants' prediction of their own score on each IAT with the average prediction of all participants for that IAT.[10] We regressed every participant's IAT *D* score onto their own prediction and this averaged prediction at the same time on level 1, and then modeled fixed effects for these random slopes on level 2. Our hypothesis was that there would be some consistency across the bias scores that would lead to a relationship between the average prediction and actual IAT scores. However, we believed that participant's unique predictions for their own score would have predictive validity and explain variance over and above these average predictions.

Results confirmed this hypothesis (see Table 5). In this simultaneous multi-level regression, there was a relationship between the mean prediction and IAT scores for every

[10] The same analyses conducted on the data of Study 1 yielded similar results.

participant, $b = .36$, $t(446.42) = 7.90$, $p < .001$. The general pattern of IAT scores was thus comparable across participants and hence the average prediction score was related to it. However, participants' unique own predictions explained variance over and above this average prediction, $b = .34$, $t(141.25) = 7.01$, $p < .001$.

Table 5: Study 2: Prediction models by condition including a prediction of participants' IAT scores by the average of all participants' predictions for that IAT.

Parameters	
Fixed effects	
Intercept (mean IAT score)	
IAT score predictions	.34***
Mean prediction	.36***
Predictions × Condition	.02
Mean prediction × condition	–.07
Random effect variances	
Intercept	.000
IAT score predictions	.028
Mean prediction	.000
Residuals	.477***
Goodness of fit	
-2 log likelihood	976.34

***$p<.001$

The dependent variable is participants' IAT score. All level-1 variables, including the dependent IAT scores, are standardized for each individual participant before they are entered in the analysis. "*Mean prediction"* represents the average of all participants' predictions for the IAT scores.

In sum then, participants did seem to make use of the same intuitions of what kind of response would be normative on an IAT. Nevertheless, their predictions for their own score explained variance over and above this general pattern. That is, any deviation of an individual participant's IAT pattern from the general pattern could be explained by participants' own unique predictions for their scores. Participants did thus seem to have unique access to their implicit associations over and above predicting normative response patterns based on naïve theories on implicit attitudes. In Study 3 we tried to address this issue more directly.

Study 3

Study 3 had two aims. One was to further investigate the difference between true introspection and predictions based on naïve theories about social norms. Recall that in Study 2 we approximated participants' sense of a normative prediction by averaging all participants' predictions. In Study 3 we decided to test this account more conservatively. In addition to predicting their own score, participants were asked to predict how they think the average participant in this study would respond. The order of their self vs. other prediction was counterbalanced across participants. If participants have introspective access into their own affective associations, then their own prediction should explain variance in their IAT over and above their idea of a normative pattern based on theories about social norms. We reasoned that this would be a particularly conservative test of unique insight, because we expected that participants would egocentrically base their idea for the average participant on their own intuitions (Krueger, 1998; Ross, Greene, & House, 1977). Any deviation in participants' prediction for the average participants as opposed their prediction for themselves, however, should be less related to their actual IAT scores. Hence participants' predictions for themselves should explain variance in their IAT scores over an above their normative predictions.

The second aim of Study 3 was to remedy methodological ambiguities of Studies 1 and 2 that could either have impeded the strength of the observed results, or allowed for alternative interpretations. First, we provided participants with a more nuanced sliding scale for the prediction of their IAT scores, rather than the 7-point Likert scale used previously. Second, we added two explicit thermometer ratings of "adults" and "regular people" to the explicit measures. Recall that the explicit thermometer ratings used in Studies 1 and 2 were absolute ratings of the five social groups (Blacks, Asians, Latinos, celebrities, and children). We chose these when comparing

implicit and explicit attitudes because the comparison groups were always White adults in the IAT. To make sure the explicit and implicit attitudes we measure were in fact equivalent in their targets, in Study 3 we also asked participants to rate their explicit feelings towards "adults" and "regular people" in addition to "Whites", thus allowing us to compute difference scores more analogous to the comparative scores obtained from the IATs. The purpose of both of these changes was simply to get more accurate results and rule out the possibility that certain results (e.g., the low implicit-explicit correlations) were methodological artifacts; we did not expect any meaningful differences in the pattern of results.

Method

Participants and Design. One hundred and twenty undergraduate students completed the study for course credit. One participant failed to understand the IAT instructions and did not complete the study. Of the final 119 participants, 77 (35.3%) were women, and 84% identified as White, with the remaining 19 participants identifying as Black (5), Latino (4), Asian (6), or mixed-ethnicities (4). Ages ranged from 18-32, with a median age of 18. The study consisted of a two-condition (predictions for own scores first vs. predictions for scores of the typical participant in the study first) between-subjects multi-level design. The previous condition assignment that was concerned with framing the meaning of implicit attitudes differently was not implemented again.

Materials. Materials were almost identical to the materials used in Study 2, with the exception of two changes and two additions.

Predictions for the average participant and condition assignment. One of the two additions in Study 3 was a prediction task for the "average participant". Participants were encouraged to imagine an average student from their university participating in this study and told that

they would be asked to predict this average student's responses to the questions, and that their prediction would be tested for accuracy. To reinforce this perspective-taking task participants were then asked to provide ratings for the average student on both all explicit thermometer ratings, as well as the IAT score predictions, counter-balancing the order of the self vs. other predictions.

IAT training. The IAT training procedure was oriented after the "true implicit attitudes" condition used in Studies 1 and 2, with one major exception. Because the threatening nature of presenting implicit attitudes as "true" (and explicit attitudes as something people "think of themselves") was not necessary and thus ethically undesirable for Study 3, all instances of the word "true" were omitted, and participants were simply asked to predict their "implicit attitudes" (instead of their "*true* implicit attitudes"). In line with this change, implicit attitudes were described as "the underlying attitude that gets triggered spontaneously and that might not be consciously known," and explicit attitudes were now described as "what you like once you've had time to think and reflect about it."

IAT score prediction. The IAT prediction task generally resembled the one used in Study 2 in the true-attitudes condition, except for the word "true", which was omitted from both the explanatory text and the prediction scale. Additionally, participants saw a sliding scale instead of seven buttons. This scale is depicted in *Figure 3*. In order to facilitate orientation, lines on the sliding scale indicated seven cut-offs that were labeled the same as the buttons in Study 2 (e.g., "… moderately more positive with WHITE than with BLACK"). As in Study 2, the scale ranged from 1.0 for an attitude that is "a lot more positive towards" the target social group to 7.0 "a lot more positive towards Whites [or adults, or regular people, respectively]." To further facilitate accurate choices, participants were told what number exactly the point to which they

dragged the slider represented (e.g., "you chose 2.5"), and asked to confirm their choice with the "Next" button. The computer registered .1-increments between a choice of 1.0 and a choice of 7.0. Participants thus had 61 choices to predict their own, and the average student's, attitude.

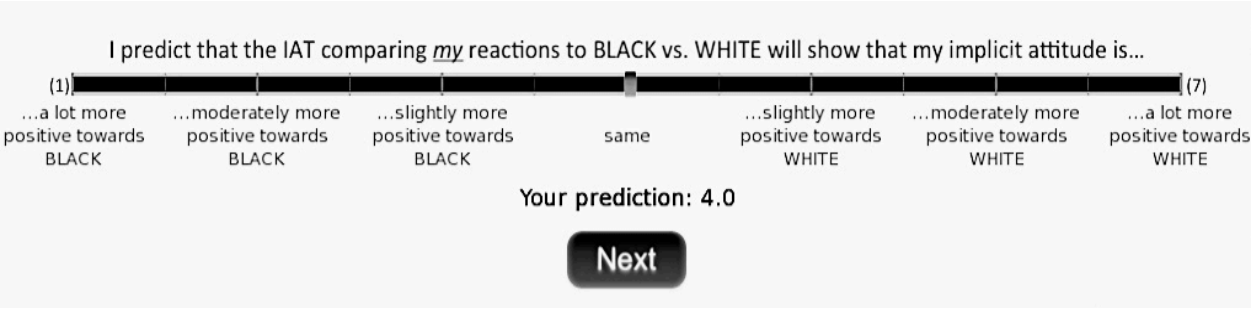

Figure 3: Prediction scale participants used to make their predictions of their IAT score (example of Black-White IAT) in Studies 3, 4, and 5. Photos used in the actual IATs were depicted above the ends of the scales on the left and right.

IAT training feedback. Participants received more precise feedback on their flower-insect and dog-cat training IATs in line with the new, more nuanced, prediction scale. Their *D* score (Greenwald et al., 2003) was converted into a numerical value between 1 and 7, with absolute *D* scores above .78 truncated to "1" or "7". Participants saw this in addition to the sentence describing their bias. Thus, a participant could see a sentence, such as, for instance, "your IAT score indicates that you have moderately more positive attitudes towards WHITE as opposed to BLACK. On the 7-point scale you used, this corresponds to a value of 5.2"). The cut-offs that were used to categorize the bias ("slightly", "moderately", or "a lot" more positive towards group x") were modified to describe intervals of equal size.[11]

[11] The original cut-offs based on the IAT webpage feedback are not equidistant and vary between .2 and .3 *D* scores (see Method of Study 1, personal communication between N. Sriram and I. Blair on July 6, 2009). To ease conversion from a *D* score to a 1-7 score that matches the bias descriptions, the *D* score cut-offs that indicate the labels were changed to describe intervals of exactly .26 *D* scores. The new cut-offs were .13 for "same" (no preference), .13-.39 for "slightly more positive towards group x", .39 - .65: "moderately more positive towards group x", and >.65: "a lot more positive towards group x". The principle change thus involves enlarging the "slight" category from .15-.35 to .13-.39; and a participant with a D score of .14 would now fall into the "slight" preference category, when they would have indicated no preference ("same") in the previous study or on the IAT webpage. The formula that was used was (*D* score*3.84615…)+4.

Additional explicit thermometer ratings. Participants were asked to rate eight groups on how warmly or coolly they, as well as the average student from their university participating in the same study, felt towards them. As in Studies 1 and 2, scales were presented on computer screens in the shape of thermometers and ranged from "0 – very coolly" to "100 – very warmly." In addition to Whites/Caucasians, Blacks/African Americans, Asians/Asian Americans, Latinos/Hispanic Americans, children, and celebrities, participants rated "adults" and "regular people (non-celebrities)" in a constraint-randomized order. That is, to avoid confusion participants rated the groups in three blocks that appeared in random order for each participant (1. ethnic groups, in random order 2. adults then children, fixed in this order, and 3. celebrities then regular people, fixed in this order).

Procedure. All participants began the experiment with an announcement that this experiment concerned their ability to predict their own scores on a computerized test accurately, as well as to predict the response of the average student from their university participating in this study accurately. They were encouraged to imagine an average participant from their university participating in this study. Participants then completed explicit thermometer ratings for themselves and for the average student, in randomized order. Next, all participants went through the IAT training procedure described above. After the IAT training participants predicted their own and the average student's score on the five IATs used in the Studies 1 and 2. They did this in the same order as they completed the explicit thermometer ratings (either self first, or average student first). After the IATs, participants again repeated all explicit thermometer ratings for themselves and the average participant, again following the same order for their self vs. other prediction. The experiment concluded with demographic information.

Results

Accuracy. We ran the same multi-level analysis conducted on Studies 1 and 2 on the data of Study 3, modeling the relationship between IAT scores and IAT score predictions on level 1, and controlling for order condition on level 2. Results are depicted in the left column of Table 6. Participants again predicted their IAT scores with considerable accuracy, $b = .59$, $t(117) = 16.26$, $p < .001$. Using the new, more nuanced 61-point predictions scales, the median correlation per participant of this skewed distribution was $r = .72$. The order condition assignment (whether or not participants predicted their own score first or the average person's score first) did not influence accuracy, $t < 1.6$, *n.s.*.

Table 6: Study 3: Prediction of own IAT score and prediction of IAT score of the average participant as predictors of actual IAT scores.

Parameters	M1 – self-prediction only	M2 – other-prediction
Fixed effects		
IAT score predictions self	.59***	.34***
IAT score predictions other		.34***
self predictions × order	–.02	–.04
other predictions × order		–.00
Random effect variances		
IAT score predictions self	.028	.009
IAT score predictions other		.033
Residuals	.505***	.452***
Goodness of fit		
-2 log likelihood	1313.49	1268.38

***$p<.001$

The dependent variables in both models are participants' IAT scores. All level-1 variables are standardized for each individual participant before they are entered in the analysis. "*Order*" represents a level-2 (between-subjects) condition assignment, one half predicted their own scores first (assigned code -1), another half predicted the score of the average participant first (coded 1).

Predictions for the average participant.

Mean pattern. Participants generally predicted the same pattern of responses for the average participant as they predicted for themselves. The mean correlation per participant between

their predictions for themselves and their predictions for the average participant was r= .73. The distribution of these correlations was highly skewed and showed a median of r=.86.[12] This mean pattern of the prediction values for self and other are plotted in *Figure 4*. We conducted a 2(prediction for self vs. prediction for other) by 5(social groups) by 2(order) condition mixed-model analysis with repeated measures on the first two factors. The principal effect of interest was an interaction of self vs. other by groups, $F(4, 468)=10.96$, $p<.001$. Simple effect contrasts showed that participants predicted that the average participant would show more bias than they themselves would show in favor of Whites as opposed to Blacks, $F(1, 117)=25.25$, $p<.001$, Asians, $F(1, 117)=10.05$, $p=.002$, and Latinos, $F(1, 117)=24.96$, $p<.001$; and more bias than they would show in favor of celebrities over regular people, $F(1, 117)=44.86$, $p<.001$, but less bias than they would show in favor of children over adults, $F(1, 117)=4.07$, $p=.046$.

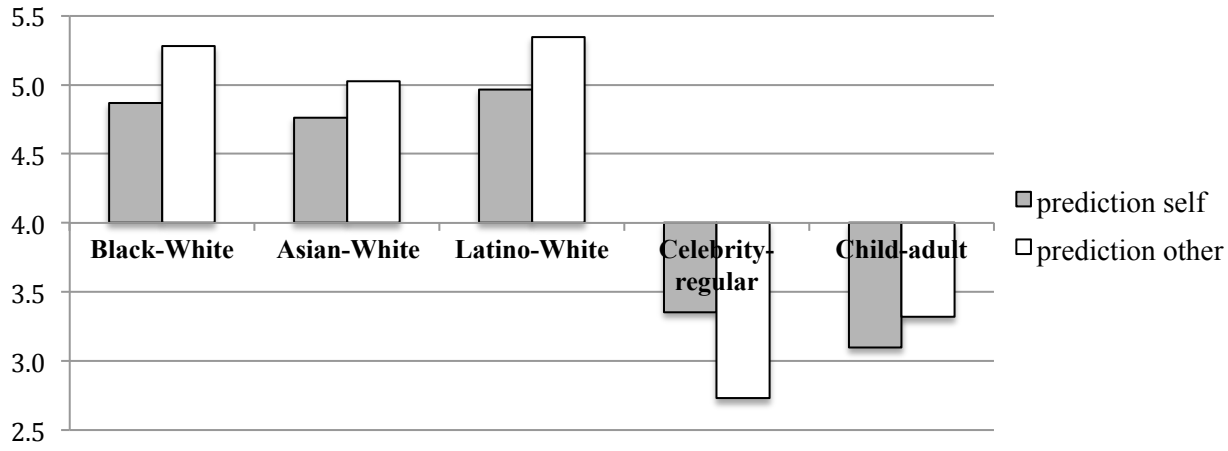

Figure 4: Study 3: Mean predictions of participants' own IAT score and predictions for IAT scores of the average participant participating in the same study. Scales range from 1-7 with scores above 4 indicating more bias in favor of the comparison group (White, regular, adult), and score below 4 indicating bias in favor of the target group (Black, Asian, Latino, celebrity, or child). All pairwise differences between predictions for self and predictions for the average participant are significant (see text).

[12] We also conducted a multi-level analysis with participants' prediction for themselves regressed on their prediction for the average participant on level 1 (both variables person-standardized), and the effect of order on this relationship analyzed on level-2. This analysis revealed that the correlation between participants' predictions for themselves and their predictions for the average participant was marginally higher when participants predicted the score for the average participant first than when they predicted their own score first, b = .06, $t(118)$ =1.85, p =.07. Since none of the other effects of theoretical interest were influenced by this order effect, it is not discussed further.

Predictions for self vs. predictions for other. In order to see whether participants had insight into their own implicit attitudes over and above the pattern they predicted for the average participant, we regressed participants' IAT scores on both their predictions for themselves and their predictions for the average participant on level 1 for each participant, and looked at how this relationship would be moderated by the order condition on level 2 across participants. Results are depicted in the right column of Table 6. Participants' predictions for the average participant was significantly related to their IAT scores, $b = .34$, $t(266.18) = 6.70$, $p < .001$. However, their prediction for their own scores explained variance over and above this relationship, $b = .34$, $t(241.72) = 7.10$, $p < .001$, indicating unique insight into their own implicit attitudes. None of these relationships were moderated by the order condition assignment, all $|t|$'s <= 1.

Explicit ratings. We again looked at the relationship between IAT scores and explicit thermometer ratings participants made for themselves. This time, each explicit rating was computed as the difference between two thermometer ratings (White minus each of the three ethnic groups, adult minus child, and regular person minus celebrity). Results are summarized in Table 7. Participants' pre-IAT thermometer ratings were moderately correlated with their IAT scores, $b = .27$, $t(118) = 5.61$, $p < .001$. However, this relationship disappeared when participants' predictions were included in the model, $b = .03$, $t(136.42) = .59$, $p = .56$.

Table 7: Study 3: The effect of explicit thermometer ratings on IAT D scores.

	M1	M2 controlling for predictions
Fixed effects		
Explicit therm. ratings	.27***	.03
IAT score predictions		.58***
Random effect variances		
Explicit therm. ratings	.111**	.074**
IAT score predictions		.020
Residuals	.655***	.454***
Goodness of fit		
-2 log likelihood	1501.44	1301.88

*p<.05 **p<.01 ***p<.001

The dependent variables in both models are participants' IAT scores. All level-1 variables are standardized for each individual participant before entered in the analysis.

We again also looked at whether or not participants would use the information about implicit attitudes and the IAT to inform their explicit thermometer ratings done after the introspection procedure, using difference scores. The relationship between these post-IAT ratings and IAT scores was significant and, as before, higher than the relationship between IAT scores and explicit ratings completed before the introspection procedure, indicating adaptation of explicit attitudes to implicit attitudes, b = .53, t(118) = 13.33, p <.001 (see Table 8). In contrast to Studies 1 and 2, this time this high relationship could not entirely be explained through participants' pre-IAT introspections. That is, although drastically reduced and thus minimal in size, the relationship between IAT scores and a post-IAT explicit thermometer ratings remained significant when controlling for participants' IAT score predictions: b = 8.25, t(121.90) =3.79, p <.001, standardized: b = .17, t(155.44) = 3.95, p <.001.

Table 8: Study 3: Explicit thermometer ratings after participants completed all IATs, regressed on their IAT D scores.

Parameters	M1	M2 – controlling for predictions
Fixed effects		
IAT *D* scores	.53***	.17***
IAT score predictions		.61***
Random effect variances		
IAT *D* scores	.052*	.074**
IAT score predictions		.062**
Residuals	.537***	.297***
Goodness of fit		
-2 log likelihood	1361.06	1112.96

*p<.05, **p<.01 ***p<.001

The dependent variables in all models are participants' explicit thermometer ratings computed as difference scores and assessed after the IAT procedure. All level-1 variables are standardized for each individual participant before they are entered in the analysis.

Discussion

Study 3 showed that participants have unique insight into their own implicit attitudes, over and above normative assumptions. The IAT score predictions participants made for themselves explained variance over and above the predictions they made for the average participant. We computed explicit attitudes slightly differently in the form of difference scores in Study 3 as opposed to the previous studies. The relationship between implicit and explicit attitudes nevertheless followed similar patterns as in the previous studies.

Study 4

The purpose of Study 4 was to examine the necessity of the IAT training procedure used in Studies 1 through 3. In the previous studies, we trained our participants on both the meaning of implicit attitudes and the IAT as a measurement instrument in order to see if introspection of implicit attitudes would be possible at all. A question left open by these findings is whether peo-

ple can differentiate between spontaneous associations and deliberate attitudes without this training.

On the one hand, previous reports on the surprise people show in reaction to their implicit-attitude results suggest that people rarely think about their spontaneous associations before taking an implicit attitude test. That is, if even the researchers who created the IAT were surprised at their own results (e.g., Banaji, 2001), and people continuously appear to be surprised at the results of implicit attitude measures (e.g., Gladwell, 2005; Nosek, 2007), then one could assume that people do not routinely think about the evaluative associations implicit attitude tests measure. These findings suggest that our participants in Studies 1 through 3 really needed the thorough explanation to be able to accurately introspect on their implicit attitudes.

On the other hand, Ranganath et al.'s (2008) findings suggest that many people can intuitively construe a difference between spontaneous and deliberate attitudes. Specifically, as described earlier, Ranganath et al (2008) asked the *same* participants to indicate both their "gut reactions" towards gay and straight people, and the "actual feelings" they experience towards the same targets "when given enough time for full consideration" (p. 388). The same participants were also asked to reflect on how their evaluative experiences developed over time in five time increments ranging from 1 "instant reaction" to 5 "given time to fully think about my feelings." Participants also completed implicit attitude measures. Important for the present considerations, results revealed that participants clearly distinguished between "gut reactions" and "instant reactions" on the one hand, and "actual feelings" or "fully considered feelings" on the other. Furthermore, a structural equation model in which "gut reactions" and "instant reactions" were grouped with implicit measures, and the fully considered attitudes were grouped separately, yielded a better model fit than a model in which all self-report measures were distinguished from

implicit attitude measures. These finding thus suggest that something as simple as a distinction between "gut reactions" and "fully-considered feelings" can be intuitive without further explanation; and that this intuitive distinction might map on to the theoretical distinction between implicit and explicit attitudes.

In Study 4, we decided to directly test the question of how much training is need to accurately predict an implicit attitude. Specifically, we manipulated both the thoroughness of explanation of the difference between implicit and explicit measures participants received, as well as experience with the IAT. Study 4 thus featured a 2 (minimal explanation vs. full explanation) by 2 (no experience vs. full experience) factorial design.

Method

Participants. One hundred and fifty-seven participants participated in this study in exchange for partial course credit. One participant responded faster than 300 milliseconds on 55% of the trials and was thus excluded in accordance with standard criteria outlined by Greenwald et al. (2003). Demographic information is available for only 154 of the remaining 156 participants. Of those, 62.33% were female, and 76.62% self-identified as White. The remaining 23.38% self-identified as Black (2), Latino (7), Asian (16, 10.39% of the sample), Native American (1), Middle-Eastern/Arab (3) or as several races/ethnicities simultaneously (7). Ages ranged from 18-32 with a median age of 19.

Design. As all previous studies, this study featured a multi-level design. The continuous relationship between participants' IAT score predictions and their actual IAT scores were modeled for each participant separately at level 1. At level 2 we modeled this relationship as a function of a 2 (minimal explanation vs. full explanation) by 2 (no experience vs. full experience) between-subjects factorial design.

Materials and Procedure. The procedure and design of Study 4 are graphically depicted in Table 9. All participants started the study by completing explicit thermometer ratings about 8 social groups as in Study 3 (White-Black, White-Asian, White-Latino, regular-celebrity, and adult-child). These scores were later used to create 5 different scores that corresponded with the IATs. They did not predict results for the average study participant; this normative prediction task was not repeated in Study 4. The condition assignments for the four conditions were implemented next. As can be seen in Table 9, the IAT training procedure can be organized into 3 steps that were exactly the same as the IAT training procedure in Study 3. They involved (1) two explanatory writing tasks on the meaning and measurement of implicit attitudes; (2) experience with predicting, completing, and receiving feedback on an insect-flower and a dog-cat IAT; and (3) reflecting on the meaning of the results of the two training IATs. After the training procedure, participants in all conditions predicted their IAT scores for the five social-group IATs in random order and then completed the actual IATs, also in random order.

Table 9: Design and procedure Study 4.

Procedure / Condition		Full explanation		Minimal Explanation	
		full experience	**no experience**	**full experience**	**no experience**
I) Explicit Thermometer Ratings		✓			
II) IAT training procedure	**Step 1**: **Explanations** and writing tasks: Implicit & explicit attitudes; IAT procedure.	✓		-- *Instead: filler writing task and prediction prompt*	
	Step 2: IAT **experience** and feedback with Insect-Flower IAT Dog-Cat IAT	✓	--	✓	--
	Step 3: **Explanatory** writing task: Reflect on your IAT results	✓ On real results	✓ On hypothetical results	--	--
III) 5 social group IATs	Score Predictions	✓ *With* reference to "gut reactions"		✓ *Without* reference to "gut reactions"	
	Actual test completions	✓			
IV) Explicit Ratings (Thermometer)		✓			
V) Demographics		✓			

Experience with the IAT (full vs. no) and explanation of the difference between implicit and explicit attitudes (full vs. minimal) was manipulated by systematically eliminating steps in the training procedure. Specifically, as can be seen in Table 9, participants in the full-explanation/full-experience condition completed all steps as participants in the previous studies. Participants in the full-explanation/no-experience condition did not complete the two practice IATs (step 2). Accordingly, step 3 for these participants involved writing a hypothetical interpretation about what their results *would* mean if they *were* to complete an insect-flower and a

dog-cat IAT. These full-explanation/no-experience participants thus predicted their IAT scores for the five social-group IATs without any experience with completing an IAT.

Participants in the minimal-explanation/full-experience condition completed only step 2 of the training procedure, but did not complete steps 1 and 3. Instead of step 1, they completed a filler task that was also titled "Do you know yourself?", but that asked participants to describe in detail what they had done on the previous afternoon (and was thus not at all related to attitudes or implicit social cognition). Participants in this condition then continued to step 2 and completed the insect-flower and dog-cat IATs and received feedback on their actual results. Hence, participants in this condition had the chance to experience completing and learning about their performance on practice IATs, but they received no specific explanation on the meaning of implicit attitudes in contrast to explicit attitudes. Their prediction was thus informed by experience with the IAT, but not by theoretical reflection about its meaning.

Lastly, participants in the minimal-explanation/no-experience condition did not complete any of the three training steps. After completing the thermometer ratings, participants in this condition completed the filler task and then went straight to predicting their scores for the social-group IATs.

In order to explain the prediction task to participants, participants in both minimal-explanation conditions were given the following prompt after the filler task, modeled after the IAT webpage's introductory portal[13], before predicting their first IAT (either insect-flower in the minimal-explanation/full-experience condition, or the social groups in the minimal-explanation/no-experience condition, see Table 9):

[13] The first two sentences are literally the same sentences visitors of the IAT webpage see, except that the words "conscious-unconscious" are replaced with "implicit-explicit" in line with the theoretical considerations presented in this paper and in line with how predictions are requested later in this study.

"This study uses a method that examines some of the divergences that may occur between people's implicit and their explicit attitudes. This new method is called the Implicit Association Test, or IAT for short. In a minute you will complete some IATs and we are interested in whether you can predict your performance on each one. Past research shows that people are actually pretty good at predicting their scores, even if they aren't entirely sure. So even if the predictions seem difficult, just try your best to be as accurate as possible."

The predictions themselves for both of these minimal-explanation conditions were also slightly modified in that they did not encourage participants to listen to their gut reactions. Instead, the screen where participants were asked to make their predictions showed the same pictures that would be used in the IATs and asked participants "if you took an IAT to measure your implicit attitude, what would it show?" The prediction scale itself was the same for participants in all conditions and the same sliding prediction scale used in Study 3 (see *Figure 3*). After completing the IATs, all participants repeated the explicit thermometer ratings, and indicated demographic information.

Results

Effects of condition assignments on accuracy. Participants' IAT scores, standardized for each participant, were regressed onto their within-participant standardized predictions for each participant separately on level 1. The resulting slopes (the prediction-actual-score relationship for each participant) were modeled as a function of training condition on level 2. Specifically, we analyzed the effect of two binary contrast-coded predictors on level 2, one for full explanation vs. minimal explanation, and one for full experience vs. no experience; and their interaction. Results are depicted in Table 10.

Table 10: Study 4: The effect of IAT score predictions and explicit thermometer ratings on IAT D scores.

Parameters	IAT score predictions only	Explicit therm. ratings only	Simultaneous Regression
Fixed effects			
IAT score predictions	.54***		.55***
Predictions × Explanation	–.02		–.02
Predictions × Experience	.00		–.01
Predictions × Explanation × Experience	.03		.01
Explicit therm. ratings		.24***	–.02
Therm. × Explanation		–.03	–.02
Therm. × Experience		.05	.03
Therm. × Explanation × Experience		.04	.04
Random effect variances			
IAT score predictions	.000	.087**	.000
Explicit therm. ratings			.058*
Residuals	.001***	.686***	.530***
Goodness of fit			
-2 log likelihood	1297.22	1997.58	1805.64

*p<.05 **p<.01 ***p<.001

The dependent variables in all models are participants' IAT *D* scores. *Therm* refers to explicit thermometer ratings computed as difference scores.

As in the previous studies, participants predicted their implicit attitudes with considerable accuracy across conditions, b=.54, t(776.0) = 17.61, p < .001. The median within-participant correlation in this study was r=.66. Surprisingly, the systematic impoverishment of the IAT training had no effect on accuracy. That is, neither explanation, b=-.02, t(776) = -.64, *n.s.*, nor experience, b =.00, t(776) =.08, *n.s*, nor the interaction of explanation and experience, b =.00, t(776) =.08, *n.s*, had any effects on the accuracy of participants' predictions.

The skew in the distribution could have affected the results. That is, because 50% of all correlations were above .65 in these skewed distributions, randomly negative, inaccurate prediction correlations were possibly overweighed when computing condition means, whereas increases in accuracy in the range above .65 between conditions might go unnoticed. For this reason, we

z-transformed all correlations by participant and then computed the mean value of those z-scores by condition. *Figure 5* graphs these mean correlations by condition, back-transformed into more easily readable Pearson-*r* correlation coefficients. As can be seen, the direction of results went in unexpected directions. If anything, participants in the minimal-explanation/no-experience condition were the most accurate in predicting their scores. A simple 2 (explanation) by 2 (experience) ANOVA was run on these z-transformed scores. None of the condition effects were significant, both main effects $F<1$, Interaction: $F(1, 152)=2.60$, $p=.11$.

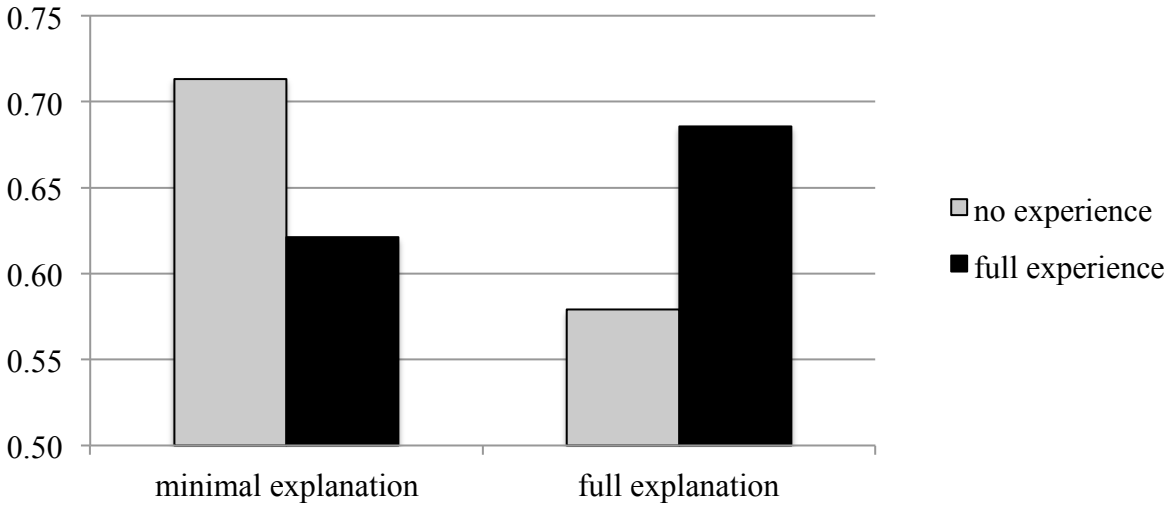

Figure 5: Study 4: Mean within-participant correlation between IAT score predictions and actual IAT scores by condition. Scores are averaged after z-transformation of each individual correlation. Mean values are then back-transformed into *r*-values for easier readability.

Relationship with explicit ratings. As in Study 3, the thermometer ratings based on difference scores showed a significant, moderate, within-participant relationship with IAT scores, $b=.24$, $t(154.00) = 5.92$, $p < .001$ (see middle and right column of Table 10). However, once controlling for participants' IAT score predictions, this relationship dropped to nil, $b=-.02$, $t(206.60) = -.48$, $p =.63$. The degree to which the predictions explained the relationship between

thermometer ratings and IAT *D* scores was constant across the four conditions, all |*t*|'s for interactions with the condition contrasts < 1, *n.s.*

As in the previous studies, participants' post-IAT explicit thermometer ratings showed adaptation to implicit attitudes. They were significantly predicted by participants' IAT scores, *b*=.47, *t*(155.00) = 13.12, *p* < .001 (see Table 11).[14] As in Study 3, this relationship was reduced, but not eliminated, when controlling for participants' IAT score predictions, *b*=.14, *t*(189.39) = 4.02, *p* < .001. The latter partial relationships did not depend on condition, all |*t*|'s for interactions with the condition contrasts < 1.6, *n.s.*

Table 11: Study 4: Post-IAT Explicit thermometer ratings regressed on participants' IAT D scores and participants' predictions of their scores.

Parameters	M1	M2
Fixed effects		
IAT *D* scores	.47***	.14***
IAT score × Explanation	–.03	.00
IAT score × Experience	.00	–.01
IAT score × Explanation × Experience	.08*	.05
IAT score predictions		.59***
Predictions × Explanation		.06
Predictions × Experience		.03
Predictions × Explanation × Experience		.01
Random effect variances		
IAT *D* scores	.051*	.055
IAT score predictions		.088***
Residuals	.577***	.321***
Goodness of fit		
-2 log likelihood	1846.84	1538.61

*p<.05, **p<.01 ***p<.001

The dependent variables in both models are participants' explicit thermometer ratings computed as difference scores and assessed after the IAT procedure.

[14] As can be seen from Table 11, this relationship was unexpectedly moderated by a three-way interaction with the explanation and experience manipulations, *b*=.08, *t*(154.00) = 5.92, *p* = .022. Looking at the nature of this three-way interaction the relationships between IAT scores and explicit thermometer ratings was higher in both the full-explanation/full-experience and the minimal-explanation/no-experience conditions, compared with the other two conditions. However, since the relationship between IAT and Thermometer ratings remained highly significant in all four conditions, the effect is contrary to expectations (and common sense), and disappears when predictions are included in the model, it will not be further discussed.

Relationship between explicit thermometer ratings and IAT score predictions. The lack of condition effects on accuracy is puzzling in many ways. In concert with the repeated finding that participants' explicit thermometer ratings were only moderately related to their IAT scores, this finding poses the question of how participants knew to differentiate between making an explicit thermometer rating and (explicitly) predicting an IAT score. In order to further investigate this process, we ran a series of additional analyses on the relationship between the thermometer ratings and IAT score predictions. The main questions of interest were (1) whether or not participants' IAT score predictions were at least related to their initial explicit feelings; (2) whether any variance in participants' predictions that was not related to their explicit ratings were indeed explained by their implicit attitudes as assessed by an IAT, as the previous results would indicate; and most importantly, (3) whether these effects would be moderated by condition assignments. In other words, did all participants in all conditions consciously indicate different attitudes when predicting their IAT scores than when completing the thermometer ratings? And does the extent to which they indicated different attitudes accurately reflect their implicit attitudes in all conditions?

Results supported these claims. They are presented in Table 12. As can be seen, there was a significant within-participant relationship between IAT score predictions and thermometer ratings, b=.47, t(151.00)=11.80, p<.001. And, although smaller in size, this relationship held when controlling for participants' IAT scores, b=.37, t(143.08)=10.30, p<.001. Importantly, this simultaneous regression also confirmed that the variance in the predictions that was left unexplained by the thermometer ratings could be explained by participants' implicit attitudes as assessed by the IATs to a substantial degree, b=.43, t(149.65)=14.97, p<.001.

Table 12: Study 4: The effect of explicit thermometer ratings on participants implicit-attitude predictions

Parameters	M1	M2
Fixed effects		
Explicit thermometer ratings	.47***	.37***
Therm. × Explanation	–.01	.00
Therm. × Experience	.04	.03
Therm. × Explanation × Experience	–.01	–.03
IAT *D* scores		.43***
IAT score × Explanation		–.01
IAT score × Experience		–.03
IAT score × Explanation × Experience		.02
Random effect variances		
Explicit thermometer ratings	.122***	.091***
IAT *D* scores		.011
Residuals	.517***	.388***
Goodness of fit		
-2 log likelihood	1815.75	1624.92

*p<.05, **p<.01 ***p<.001

The dependent variables in all models are participants' IAT score predictions computed as difference scores and assessed after the IAT procedure.

Crucially, results indicated that none of these effects were moderated by condition, all $|t|$'s <=1.09. That is, all participants in all conditions consciously indicated different attitudes when they predicted their IAT scores than when they completed the thermometer ratings, even those in the minimal-explanation/no-experience condition. And in all conditions, these differences accurately reflected participants' actual IAT *D* scores (and although no difference reached significance, looking at simple slopes, participants in the minimal-explanation/no-experience condition in fact if anything made predictions that were most different from thermometer ratings, and most in line with their implicit attitudes as assessed by the IATs, compared to the other conditions).

Discussion

The purpose of Study 4 was to assess whether people would be able to accurately introspect on their implicit attitudes even without substantial explanation of the differences between implicit and explicit attitudes, and without experience with an implicit attitude measure. Results indicated that neither of these factors are necessary conditions for accurate introspection. Participants were equally accurate (if anything slightly more accurate) in predicting their implicit attitudes even when they received minimal explanation and had no direct experience in the same study with completing an implicit attitude measure, as they were with full training.

One might take these results to indicate that people have substantial understanding of the concepts "implicit" and "explicit" attitudes. If this explanation were correct, then telling participants that we are interested in "implicit-explicit divergences," as the prompt did, would be enough to make them accurately predict their implicit attitudes. However, given that our participants were undergraduate freshmen (and the topic was positively not covered in their intro-psych lectures), and that the term "implicit attitudes" is rarely used colloquially, this explanation seems somewhat unlikely. Unfortunately, we did not ask participants whether or not they had heard about the concept of implicit attitude or completed an implicit attitude measure in another study or on the Internet before participating in our study. Yet, even if a sizable proportion had, this would not explain why participants in the minimal-explanation/no-experience condition on average ended up giving equally accurate (if anything more accurate) predictions as participants who received full training.

If we assume that participants did not come into our study as experts in the field of attitudes, the question remains, how did they know to predict their attitudes accurately, and how were they motivated to predict implicit attitudes that were different from the explicit attitudes

they had predicted just moments prior? To understand this point, it might be helpful to take a deeper look at the differences between the two measures. From the perspective of participants in the minimal-explanation/no-experience condition: How did predicting an implicit attitude differ from indicating an explicit attitude? The four most striking differences are discussed below:

First, as already mentioned, a prompt announced that we were interested in "divergences" in participants' attitudes. Participants were thus sanctioned to make indications for their predictions that "diverged" from the attitudes they had just indicated. This might be a crucial point, but it does not yet explain how participants knew to pay attention to their gut reactions for the second rating more than the first.

Second, we announced, and repeated in every prediction, that we would in fact accurately measure participants' implicit attitudes later – they were predicting the outcome of a computerized task. This announcement presumably functioned as a "bogus pipeline" (Jones & Sigall, 1971). That is, any self-presentational and other distortions participants make when indicating their explicit attitudes are obviously useless if there is certainty that the "truth" will be revealed. As mentioned earlier, self-presentational concerns have been shown to be a reliable source of reduction of implicit-explicit correspondence (Nosek, 2005). On the other hand, the effect of self-presentational concern found is not particularly large (Hofmann et al., 2005a, 2005b; Nosek, 2005). Additionally, simply telling participants that their attitudes will be measured with a task has had no known effect on implicit-explicit correlations so far. For instance, although this is not discussed in a paper to our knowledge, visitors on the IAT webpage know when they enter the portal that their implicit attitudes will be measured before any data is collected (the first prompt is in fact almost synonymous with the prompt given to our participants in the minimal-explanation conditions) – yet many publications on implicit-explicit relations that use website

data (e.g., Nosek, 2005; Nosek & Hanson, 2008) still report implicit-explicit relationships that are substantially lower than the accuracy data reported here (i.e., Nosek, 2007; Nosek & Hansen, 2008).[15] Hence, an announcement that attitudes will be measured could only have the effect reported here in concert with other factors.

Third, implicit predictions were made with reference to specific faces. As mentioned in the introduction, one piece of information that could presumably influence an explicit attitude more than an implicit attitude is bringing specific exemplars to mind. A person might indicate more positive explicit attitudes than an IAT would show towards African Americans, for instance, because the person is thinking of admired Black individuals that they actually feel positively about when they make their explicit rating (e.g., Barak Obama, Martin Luther King, etc.). It is possible that participants have a strong sense that their feelings towards a category presented as a series of non-smiling faces of strangers is different than their feelings that they brought to mind when asked to make a more abstract rating.

Fourth and last, participants predicted their IAT scores in a relative fashion ("more positive towards category X than towards category Y"). Explicit thermometer ratings were completed individually and we computed difference scores afterwards. Participants might always decide on different attitudes when social groups are considered in isolation than when they are asked to consider them conjointly.

In sum, awareness that "different" attitudes that "diverge" will be measured, awareness that scores will be revealed regardless, rating specific unknown exemplars as opposed to abstract category labels, and making relative as opposed to absolute ratings, could all have contributed to

[15] Implicit-explicit relationships are analyzed between-subjects in those papers (Nosek & Hansen, 2008) or as crossed (i.e., scores are nested under participants and attitude targets at the same time in Nosek 2007). As discussed later in this paper, this method of analysis could have technically underestimated participants' awareness of the pattern of their own attitudes. We will return to this point later in a final section on the method of analysis.

participants' awareness that a prediction of an implicit score should be different than the explicit thermometer ratings they completed just several minutes prior. What remains surprising, however, is that the differences between thermometer ratings and implicit score predictions were in fact in line with participants' implicit attitudes as later assessed by an IAT. That is, what Study 4 shows more overwhelmingly than any of the previous studies, is that people really are highly aware of their implicit attitudes; and that one does not have to dig very deep to reveal them. Future research is clearly needed to explain exactly how people construe implicit associations in contrast to other propositional explicit attitudes, and why so many standardly used self-report measures only capture the latter.

Study 5

The purpose of Study 5 was to show that the factors that influence the relationship between implicit and explicit attitudes have no bearing on a person's ability to introspect their implicit attitudes. Specifically, as described earlier we assume that the main difference between implicit attitudes and explicit attitudes lies in their representation. Implicit attitudes are assumed to be based on spontaneous, mostly uncontrollable, associations that subjectively manifest themselves in the form of "gut reactions." Explicit "feelings," on the other hand, are supposed to be the result of a propositional process by which a person considers a variety of propositions before deciding on their "feeling" towards an attitude object. We reasoned that if this theoretical framing is correct, then an experimental manipulation that encourages people to consider more propositional information should decrease the correspondence of implicit and explicit attitudes. On the other hand, we were curious to see whether introspections of implicit associations would also be affected by such a manipulation.

Previous research has shown that deliberating reasons for one's attitudes decreases correspondence between these explicit attitudes and behavior (e.g., Wilson, Dunn, Bybee, Hyman, & Rotondo, 1984; see Wilson, Dunn, Kraft, & Lisle, 1989). It has also shown that deliberating reasons for an attitude decreases correspondence between explicit attitudes and implicit attitudes (Gawronski & LeBel, 2008). In an attempt to replicate the latter findings, we decided to implement a propositions manipulation. Specifically, we gave one group of participants instructions to consider reasons for their feelings before completing their thermometer ratings. We expected that this group's explicit thermometer ratings would be correlated less with IAT scores than in previous studies. Another group of participants was instructed to try not to think too much about their attitudes and just trust their initial thoughts. In line with our theorizing, we predicted that this group's explicit attitudes would correspond more strongly to their implicit attitudes than in the previous studies. A third control group was not given specific instructions before indicating their attitudes. These participants were expected to show moderate implicit-explicit attitude relations, comparable to those found in the previous studies.

Importantly, we expected participants' introspections of their implicit attitudes to remain unaffected by these manipulations. That is, we reasoned, if participants are truly aware of their spontaneous evaluative associations, then they should know to leave an explicit report on those associations uncontaminated by additional propositional information.

Method

Participants and design. One hundred and twenty-two CU Boulder undergraduates participated in the study in exchange for partial course credit. Six participants who initially participated experienced computer errors and were excluded. Five additional participants responded faster than 300 ms on more than 10% of the trials and were deleted in line with standard criteria

(Greenwald et al., 2003). Demographic information was available for 102 of the remaining 111 participants. Of those, 45.1% indicated being male, and 84.3% indicated being White. The remaining 15.7% indicated being Black (2), Latino (3), Asian (2), Native American (1), Middle-Eastern/Arab (2), or mixed ethnicities (6). Ages ranged from 18 to 28, with a median age of 19.

As all previous studies, Study 5 consisted of a multi-level design. The continuous relationships between IAT scores on the one hand, and explicit thermometer ratings and IAT score predictions on the other hand, were modeled for each participant separately on level one. The resulting slopes of these relationships were modeled as a function of three explicit thermometer-rating conditions (reasons, control, intuitive) on level 2.

Materials and procedure

Explicit thermometer ratings and condition assignments. All participants started the study with the explicit thermometer ratings of the eight social groups used in Studies 3 and 4. As previously, we later used these eight ratings to create five difference scores that corresponded with the IATs (White-Black, White-Asian, White-Latino, regular-celebrity, and adult-child). In order to implement the propositions manipulation, instructions for these explicit thermometer ratings were presented in three different ways. Specifically, participants in the "reasons" condition were told that we were interested in the reasons "people feel the way they do" about social groups. They were then asked to consider and indicate three reasons for their feelings before every single thermometer rating in a free-response mode in three separate text boxes. All participants in this condition completed all three text boxes for all groups they rated. The thermometer ratings themselves also started with a sentence reminding participants to think about the reasons they just indicated when entering their ratings.

Participants in the "intuitive" condition were told that we were particularly interested in their immediate and spontaneous feelings. Each individual thermometer rating was then also accompanied by a sentence reminding participants to pay attention to their "immediate and spontaneous feelings." Participants in the control condition were simply asked to indicate their feelings towards the eight social groups with no additional specification.

After the thermometer ratings participants in all conditions went through the same IAT training procedure as participants in Study 3, and the full-explanation/full-experience condition in Study 4. They then predicted their IAT scores towards the social groups, completed the IATs, and repeated the thermometer ratings (without varying instructions by condition). The study concluded with demographic information.

Results

Mean pattern of explicit ratings. The mean computed difference scores that correspond with the IATs are shown in Table 13. Contrary to hypotheses, none of the ratings differed significantly by condition, all omnibus F's <1.70, all p's > .19.

Table 13: Study 5: Mean thermometer ratings (computed as difference scores) per condition.

	Condition		
Difference Scores:	Reasons	Control	Intuitive
White – Asian	12.24	15.26	8.92
White – Black	7.47	10.23	6.95
Regular People – Celebrities	23.5	28.1	17.97
Adults – Children	-7.74	-6.44	-6.03
White – Latino	7.88	15.9	13.79

Relationships between thermometer ratings, IAT score predictions, and IAT scores, within-subjects. The explicit thermometer rating difference scores were standardized for each participant and regressed onto participants' IAT scores (also standardized) individually on level 1. The resulting slopes were modeled as a function of two contrast-coded predictors for condition

on level 2: One comparing the reasons condition (=-1) to the intuitive condition (=1); and another one comparing the control condition (=-2) to the other two conditions (both =1). Results are presented in the left-most column of Table 14. There was a moderate average within-participant relationship across conditions comparable to the relationship in previous studies, b=.24, t(552.0) = 5.68, p<.001. However, contrary to hypotheses, this relationship remained completely unaffected by the condition assignments, all possible orthogonal contrast codes: t's < 1, omnibus effect of the condition-by-thermometer interaction, F(2, 552.0) = .03.

Table 14: Study 5: The effect of predictions and explicit thermometer ratings on actual IAT D scores

Parameters	Explicit therm. ratings only	IAT score predictions only	Simultaneous Regression
Fixed effects			
Explicit therm. ratings	.24***		–.01
Therm. × Contrast 1: reasons vs. intuitive	–.01		.03
Therm. × Contrast 2: control vs. other	–.01		.00
IAT score predictions		.50***	.51***
Predictions × Contrast 1		.01	–.01
Predictions × Contrast 2		–.00	–.01
Random effect variances			
Explicit therm. ratings	.000		
IAT score predictions	.001	.042	.045
Residuals	.760***	.569***	.570***
Goodness of fit			
-2 log likelihood	1433.34	1301.65	1314.94

*p<.05 **p<.01 ***p<.001

The dependent variables in all models are participants' IAT scores. *Contrast 1* refers to a contrast-coded predictor of condition assignment that assigned the following codes: reasons = -1, intuitive =1, and control=0. *Contrast 2* assigned the following codes: reasons = 1, intuitive =1, and control=-2.

Within-subject accuracy of predictions was comparable to previous studies, b=.50, t(107.0) = 12.29, p<.001, and as expected, also completely unaffected by condition, t's<.1, omnibus condition-by-prediction interaction, F(2, 107.0) = .02 (see middle column, Table 14). In order to see whether or not the unique thermometer-IAT relationship by condition would come

out when controlling for participants' IAT-score predictions, we also regressed standardized scores of participants' IAT scores on both IAT score predictions and difference thermometer scores simultaneously. Results are presented in the right-most column of Table 14. As in the previous studies, there was no unique thermometer-IAT relationship when controlling for participants' predictions, b= -.01, $t(134.1) = .17$. And this unique (non-)relationship, too, was entirely unaffected by condition, all t's <1, omnibus interactions with both thermometer and prediction: $F(2, 137, 56) <.19$. That is, as before, all relationships between the thermometer ratings and IAT scores could be entirely explained through participants' awareness of their IAT scores; and importantly, this was not more the case in any one condition as opposed to the others.

The results just presented are puzzling. Importantly, they fail to replicate the well-documented effect that implicit-explicit correlations increase with instructions to make more intuitive judgments (Smith & Nosek, 2011; Ranganath et al., 2008). They also fail to support the theoretical idea that considering propositional information should decrease implicit-explicit correspondence (Gawronski & LeBel, 2008).

One reason for this failure of replication could be the method with which we analyze these data. Specifically, the previous findings cited all concern the magnitude of implicit-explicit correlations *between-subjects* (Smith & Nosek, 2011; Ranganath et al., 2008; Gawronski & LeBel, 2008). We analyzed our data within-subjects, because we were curious to see whether or not participants could tell how their implicit associations would relate to each other, regardless of their estimations of how their implicit attitude relate to those of other people. Still, given the lack of replication of established effects, we decided to look at our hypotheses on implicit-explicit relations in a between-subject analysis.

Between-subjects analyses. We first decided to look at between-subject correlation of the variables for each target group pair separately (i.e., Black-White, Latino-White, Asian-White, child-adult, and celebrity-regular), within each condition. The values represented in Table 15 are averages across the five target-pair correlations. For easier comparison they are depicted next to the average correlations per condition that result from computing within-subject correlations for each participant and then averaging them within each condition, as presented earlier. As can be seen, the pattern of the between-subject correlations does in fact go in the expected direction: Implicit-explicit correlations were higher when participants were instructed to make their thermometer ratings based on spontaneous reactions, than when they were told to complete these thermometer ratings based on reasons they had just considered. The size of these correlations is furthermore comparable to results presented by Hofmann et al. (2005a) in their meta-analysis. Hofmann et al. report that the average implicit-explicit correlation across the studies they investigated for affectively based explicit measures (assessed as relative ratings between two groups) was .30, while the average implicit-explicit relationship for more cognitively based explicit measures (also assessed as relative measures) was .21. Our average between-subject correlations are lower, but Hofmann et al.'s (2005a) difference corresponds to the difference in the between-subject correlations between the intuitive and the reasons condition in our study. As the right-most columns of Table 15 show, this pattern was different than the pattern observed when the relationships were computed for each participant within-subjects across IATs.

Table 15: Average correlations by condition.

	between-subjects		within-subjects	
Condition	Therm.-IAT	Predictions-IAT	Therm.-IAT	Predictions-IAT
Reasons	.18	.25	.24	.49
Control	.21	.23	.25	.51
Intuitive	.27	.28	.22	.51

Correlations are computed once between subjects (left, average correlation of 5 correlations per condition computed across participants in each condition), and once within-subjects (right, average of as many individual correlations as there are participants per condition, across five IATs).

Predicted *partial* simple slopes by condition that result of a simultaneous regression with both predictions and thermometer ratings as simultaneous predictors, made this pattern even clearer: In these analyses, we estimated separate simultaneous regression models for each target pair between-subjects, and then averaged the slopes across the five target pairs. Results are depicted in Table 16, in the two left columns (the right columns represent the same values estimated within-subjects for comparison). In these analyses, participants' thermometer ratings and predictions had about equal unique relationships with IAT scores in the intuitive condition. When participants were asked to consider reasons for their attitudes, however, participants' thermometer ratings had no relationship with IAT scores over and above their predictions of these scores. Again, this pattern differed from observations made within-subjects.

Table 16: Study 5: Average partial regression slopes by condition

	Between-subjects		Within-subjects	
Condition	Therm.-IAT	Predictions-IAT	Therm.-IAT	Predictions-IAT
Reasons	.06	.21	–.04	.51
Control	.11	.14	–.02	.52
Intuitive	.20	.20	.02	.50

As in Table 15, partial regression slopes are computed once between subjects (left) and once within-subjects (right)

Nevertheless, none of these differences reached conventional levels of significance. Specifically, in order to analyze these effects, we standardized all IAT scores and Thermometer difference scores, this time across participants per IAT. We then regressed each standardized IAT score on a standardized thermometer difference score (and later the predictions) for the same tar-

get pairs between-subjects, including two contrast-coded condition predictors, and their interaction, on level 1. We thus estimated 5 regressions on level 1, one between-subject regression for each target pair. At level 2 we looked at the average of these effects across the five target pairs.[16] Results are presented in Table 17.

*Table 17: Study 5: The effects of explicit thermometer ratings and score predictions on IAT scores, analyzed **between** subjects.*

Parameters	Explicit therm. ratings only	IAT score predictions only	Simultaneous Regression
Fixed effects			
Intercept	.00	.01	.00
Contrast 1: reasons vs. intuitive	–.02	–.01	–.01
Contrast 2: control vs. other	–.01	–.00	–.00
Explicit therm. ratings	.23*		.12†
Therm. × Contrast 1	.08		.07
Therm. × Contrast 2	.02		.00
IAT score predictions		.26**	.18*
Predictions × Contrast 1		.04	–.00
Predictions × Contrast 2		.02	.02
Random effect variances			
Intercept	.000	.000	.000
Contrast 1: reasons vs. intuitive	.013	.014	.013
Contrast 2: control vs. other	.000	.000	.000
Explicit therm. ratings	.007		.006
Therm. × Contrast 1	.002		.000
Therm. × Contrast 2	.000		.000
IAT score predictions		.004	.002
Predictions × Contrast 1		.008	.009
Predictions × Contrast 2		.000	.000
Residuals	.933***	.921***	.912***
Goodness of fit			
-2 log likelihood	1563.99	1557.75	1563.60

†p =.096 *p<.05 **p<.01 ***p<.001

The dependent variables in all models are participants' IAT scores. *Contrast 1* refers to a contrast-coded predictor of condition assignment that assigned the following codes: reasons = -1, intuitive =1, and control=0. *Contrast 2* assigned the following codes: reasons = 1, intuitive =1, and control=-2.

[16] Note that condition effects are nested under IAT on level 1 in these analyses. In the within-subject analyses condition effects were analyzed at level 2. Accordingly, Table 17 contains intercepts, main effects of condition, as well as estimates of random effects for these effects.

Both regression analyses (simple and simultaneous) showed that the reported differences in the relationships between IAT scores and thermometer ratings by condition were not significant, interaction of thermometer with a contrast code comparing the reasons condition (=-1) with the intuitive condition (=1) in a simple analysis: b=.08, t(4.12) = 1.36, p=.24; same interaction run with predictions in the model in a simultaneous regression: b=.07, t(6.33) = .99, p=.36.

In sum, there was limited indication that considering reasons for their attitudes reduced our participants' explicit attitudes' correspondence with their implicit attitudes. Whereas previous patterns were somewhat replicated between-subjects, no analysis showed significant effects of the manipulations. The accuracy of participants' predictions also remained unaffected by the propositions manipulation.

Discussion

The purpose of Study 5 was to show that the amount of propositional information a person considers when making an explicit attitude rating influences this rating's correspondence with implicit attitude measures. A manipulation aimed at increasing the amount of propositional information a person considers yielded no support. The correspondence of explicit thermometer ratings and implicit attitudes was statistically indistinguishable for participants in the different conditions. Participants further continued to predict the patterns of their implicit attitudes accurately.

When analyzed as relationships between participants, the relationships between implicit and explicit attitudes per condition did align with predicted patterns somewhat, but still failed to reach significance. The comparable sizes of the correlations with values reported in the meta-analysis by Hofmann et al. (2005a) might suggest that our analyses might have been statistically underpowered. However, using 5 different IATs and more than 35 participants per condition

should be a somewhat respectable amount to meet significance levels for a medium-size effect. Future research is clearly needed to see if a reasons manipulation modeled after Wilson et al. (1984, 1989) will decrease implicit-explicit attitude correspondence.

Another interesting point raised by the previous analyses is the difference in results when we analyzed our data within-subjects as opposed to between-subjects. That is, especially the accuracy of participants' predictions of their IAT scores seemed strikingly different depending on the method of analysis. We turn to this issue next, in a last analysis across all five studies.

Telling if something tastes sweeter to you than to others –

Within vs. between-subject assessment of accuracy

In all previous studies, we analyzed almost all relationships between participants' IAT score predictions and their actual IAT scores within-subjects, across five IATs. Given that most research on implicit-explicit attitude correspondence is computed between-subjects, however, we were curious to see how participants' predictions of their IAT scores would fare in predicting their actual scores between-subjects as well. Note that, as described in the beginning, this analysis answers a slightly different question than the analyses presented previously. Specifically, the work presented in this paper so far was mainly concerned with the question of whether people would be able to tell that they have more or less positive associations with some social categories as opposed to others. A within-subject analysis answers exactly that question. Results showed that participants were impressively accurate at predicting the pattern of their implicit attitudes towards 8 different target groups on five different IATs.

A between-subject analysis would answer an additional, different question. Here, we are looking not only at whether a person can tell that they have more or less positive associations with one social category as opposed to another, but whether or not they know if their associa-

tions are more or less positive than those of other people, and exactly to what degree. We decided to look at all relationships between IAT score predictions and actual IAT scores between-subjects compared to within-subjects. Results are presented in Table 18 for all five studies, collapsing across conditions.[17] Because the distribution of the within-subject correlations, especially those between the predictions and the IAT scores, were highly skewed, we also computed means of z-transformed (and hence un-skewed) values and back-transformed the resulting means into more easily readable r-scores. The results are presented in panel b) of Table 18.

Table 18: Average correlations between IAT D scores, and both IAT score predictions and Thermometer ratings by study.

a) Raw correlations

	within-subjects		between-subjects	
	Pred.-IAT	Therm.-IAT	Pred.-IAT	Therm.-IAT
Study 1 (N=65)	.51	–.01	.33	.23
Study 2 (N=90)	.53	.20	.39	.23
Study 3 (N=119)	.59	.27	.28	.21
Study 4 (N=156)	.53	.24	.31	.21
Study 5 (N=111)	.50	.24	.26	.23
Average across studies	.53	.19 (.24)	.31	.22

[17] Condition effects (and lack thereof) were mostly replicated in the between-subject analyses. Specifically, as in the within-subject analyses, there were no condition effects in Studies 3, 4, and 5; and there was a marginal trend for predictions to be more accurate in the attitudes condition as opposed to the associations conditions in Study 1, b=.09, t(321.0) = 1.73, p=.09. The only difference was a trend for less accuracy in the attitudes condition as opposed to the associations condition in Study 2, b=-.06, t(442.0) = -1.40, p=.16. Looking at the individual IATs, this effect was driven entirely by the Black-White IAT. That is, participants were more accurate in predicting their Black-White IAT score between-subjects, when they thought of it merely representing a cultural association, r=.52 than when they thought about it as representing their true implicit attitude, r=.00; but only in Study 2 and not in Study 1. This pattern was not replicated for any of the other minority groups either. It should thus be treated with caution.

b) Mean of un-skewed correlations, back-transformed

	within-subjects		between-subjects	
	Pred.	Therm.	Pred.	Therm.
Study 1 (N=65)	.63	–.02	.34	.23
Study 2 (N=90)	.65	.25	.39	.24
Study 3 (N=119)	.72	.37	.28	.21
Study 4 (N=156)	.65	.33	.31	.21
Study 5 (N=111)	.62	.30	.26	.24
Average across studies	.65	.25 (.31)	.32	.23

Within-subject correlations are computed for each participant and then averaged across as many participants as each study contains. Between-subject correlations are computed per IAT and then always averaged across 5 IATs.
Panel a) shows results of averaging raw correlations (of each participant or of each IAT).
In panel b) all individual correlations per subject or per IAT are z-transformed and then averaged. The values depicted are back-transformations of these average z-scores into correlation values. Between-subject correlations are based on only 5 values and in a moderate range. Hence their distributions are not skewed and transformation leads to similar results.
Thermometer ratings are simple scores in Studies 1 and 2, but reverse scored for easier comparability. Thermometer ratings in Studies 3-5 are computed as difference scores comparable to the IATs.

As can be seen, the prediction-IAT scores relationship was not nearly as high when computed between-subjects, as it was when computed within-subjects. Additionally, this difference was not observable to the same degree for the thermometer ratings. The last point supports the interpretation that the increase in relationship for the predictions is not a methodological artifact, but much rather a theoretical point. Specifically, while participants were impressively accurate in predicting how their IAT scores would relate to each other, they had much more limited insight into how their scores would relate to those of other people.

Another way to look at this difference comes from an examination of the average IAT score predictions and the average actual IAT scores. They are depicted in *Figure 6*, across studies. At first glance, one can again see the accuracy with which participants predicted the pattern of their IAT results. When taking labeling conventions into consideration, however, this image changes. Specifically, recall that on the 1-7 prediction scales participants used, an attitude was

labeled as a "slight" preference for one group over another for the scale points 3 and 5. The scale points 2 and 6 represented a "moderate" preference, and the ends of the scales, 1 and 7, represented a "strong" preference. Taking these anchors into consideration, one can see that participants mostly perceived their own implicit preferences to be less than "slight."

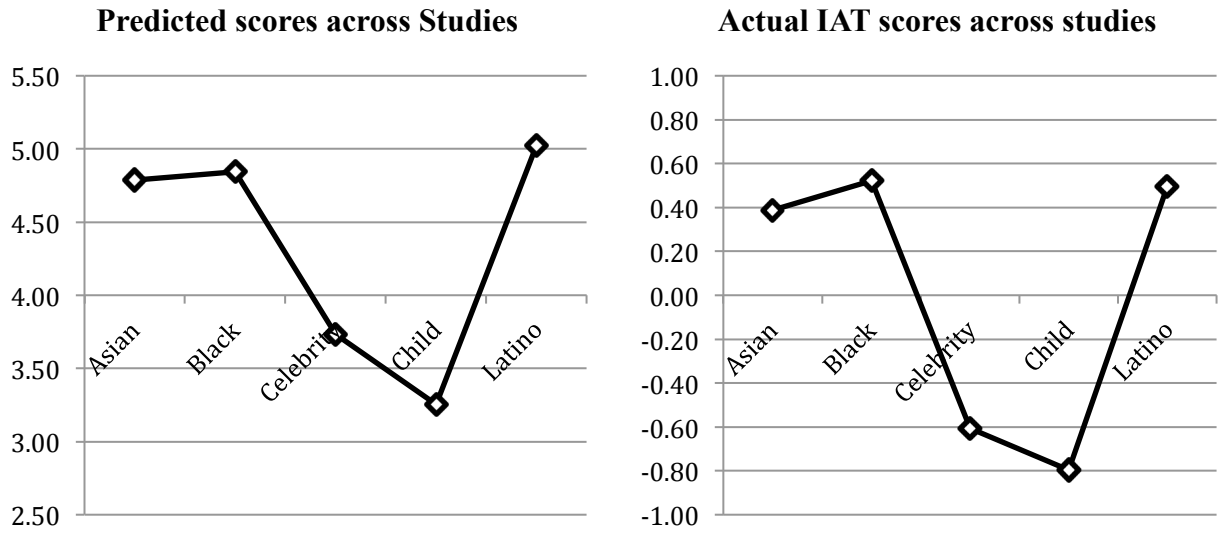

Figure 6: Average IAT score predictions (1-7 scale) and average actual IAT scores (*D* scores) across studies. Studies are weighted by size (i.e., each participant contributes equally to these mean scores).

In contrast, according to IAT scoring conventions described earlier (Personal communication from N. Sriram to I. Blair on July 6, 2009), the conventions for a "slight", "moderate", or "strong" preference in *D* scores lie at .15, .35, and .65. The right side of *Figure 6* indicates that, according to these conventions, participants on average had "moderate" to sometimes even "strong" preferences.

One could thus argue that participants in fact underestimated their biases. Note, however, that while the IAT cut-offs are oriented after statistical conventions for effect sizes, they have no absolute value in social reality. That is, whether or not the negative gut reaction a person feels in reaction to a social category is "slight", "moderate", or "strong," is entirely subjective and has no

objective truth value attached to it. We would thus argue that one could never examine accuracy of self-insight with between-subjects analyses, since different subjective experience, and labeling preferences will inevitably skew them. In fact, the inevitable subjectivity of psychological experience might make a within-subject analysis the only way one can study introspective accuracy, awareness, and self-insight.

General Discussion

The purpose of the current set of studies was to investigate a common assumption that implicit attitudes cannot be introspected. In line with the APE model (Gawronski & Bodenhausen, 2006) and in contrast to a common assumption, we believed that people would be reasonably accurate at predicting their scores on future IATs they had not yet completed. Results from five studies supported this hypothesis. Participants predicted scores on five IATs with considerable accuracy across three different prediction methods. A manipulation aimed at testing whether self-presentational concerns would impede people's ability to accurately introspect and report their implicit attitudes yielded no support. Participants were equally accurate at predicting their scores when those were portrayed as revealing unpleasant truths about the participants, as when they were portrayed as revealing non-threatening information. The only exception to this finding was a lower between-subject correlation for a Black-White IAT in the attitudes condition as opposed to the associations condition in Study 2 (see Footnote 17). Because American Black-White relations might be the most over-studied exemplar of intergroup attitudes in general and implicit attitudes in particular, this might justify many researchers' intuitions that people would be hesitant to admit their implicit attitudes when these might reveal unpleasant truths about them. However, given that this effect did not show in Study 1 (the pattern went in the opposite direction in that Study even for the Black-White IAT), that it

did not show for any other racial and ethnic groups, and accordingly, that it did not show in the within-subject analysis, it should be treated with caution. Nevertheless, many Americans might be particularly hesitant to admit their negative attitudes towards African Americans, more than towards other social groups. Accordingly, much theorizing about implicit attitudes in general might be over-weighted by this one particular example that might be more the result of a specific socio-historical development than a psychological truth about attitudes.

We also conducted an analysis aimed at investigating whether participants had unique insight into their own reactions or simply reported attitudes based on theories they might hold about social norms. We investigated this question with two different approximations of participants' normative theories. Both analyses supported the notion that participants have unique insight into their own implicit attitudes over and above normative ideas.

Surprisingly, introspective accuracy was also independent of the training participants received in the same study. In Study 4, participants were equally accurate when they were simply asked to predict their implicit attitudes that could "diverge" from their explicit attitudes (without further explanation of the terms), as they were when they had been thoroughly explained the constructs and experienced their measurement first. This last finding suggests that implicit evaluative associations are even more readily accessible to people than the first three studies would suggest.

A last analysis showed that while participants were impressively accurate at predicting the pattern of their implicit attitudes, they had somewhat limited insight into how their attitudes would relate to those of other people. A look at the mean predictions furthermore revealed that across studies participants classified their own implicit biases as "slight" at best, although IAT conventions would classify them as moderate. They also estimated their own biases to be

weaker than the biases of the average study participant in Study 3. It appears that awareness of one's intergroup biases does not preclude engaging in self-serving interpretations. At the same time that participants interpreted their biases to be "slight" to barely noticeable, they nevertheless showed considerable awareness of very nuanced variations in their associations with eight different social groups.

One possible alternative interpretation of our accuracy results could be that participants willingly changed their IAT scores so that they would represent the implicit attitudes they predicted. We believe this interpretation to be unlikely for the following reasons. There is considerable evidence that it is hard to impossible to fake an IAT result (which is in fact one of the reasons this attitude measure has been as attractive to so many researchers, Asendorpf, Banse, & Mücke, 2002; Banse, Seise, & Zerbes, 2002; Egloff and Schmukle, 2002; Kim, 2003; see Greenwald et al., 2009). So far, the only apparently successful strategy found is willingly slowing down on one IAT block as opposed to another (Greenwald et al., 2009). However, Greenwald and colleagues report that, first, few participants spontaneously discover this strategy. Second, in our Study 4 results were equally accurate when participants were not even explained how their implicit attitudes would be measured (making a plan to use such a strategy during the prediction task highly unlikely). Lastly, the result of such a strategy would likely be an exaggerated shift in the results, not a change in attitudes as nuanced as the subtle variations between the five different social-group IATs that our participants accurately predicted. We thus believe it unlikely that our high accuracy results stem from participants being able to willfully produce specific results on their IATs.

Relationships between implicit attitudes as assessed by the IAT, and explicit attitudes assessed by a "thermometer" rating, followed a different pattern. First, any relationship observed

between explicit thermometer ratings completed before the IATs and actual IAT scores were low if present at all. Furthermore, in cases where we did find a relationship, this relationship could be entirely explained through participants' awareness of their IAT scores. Recall the example of a White American thinking about the category "Black" in the beginning of this paper. We argued that this White subject could be entirely aware of their implicit reactions, but might nevertheless decide on a different explicit attitude after considering other propositions. The initial implicit association would, if we consider awareness possible, be one proposition that factors into this person's explicit attitudes, but only one of many. The current data strongly support that explanation. The relationship between implicit attitudes and explicit ratings was always lower than the relationship between implicit attitudes and implicit-attitude predictions, and any relationship between initial thermometer ratings and IAT scores could be explained through participants' awareness. We believe this finding supports the interpretation that participants' implicit attitudes were just one of many propositions participants considered in their formation of their explicit judgment.

Nevertheless, contrary to this theorizing, a manipulation aimed at heightening the amount of propositions participants use to decide on their explicit attitudes yielded no results. More research is needed to further investigate how participants decide on their explicit attitudes, and under what circumstances those will deviate more or less from their implicit associations.

Interestingly, a re-assessment of explicit attitudes after participants completed the IAT predictions and IATs indicated that participants adapted their explicit attitudes to their implicit attitudes. We believe that this finding is also in line with the APE model (Gawronski & Bodenhausen, 2006). Specifically, the APE model asserts that explicit attitudes are the result of a propositional process by which a person considers all propositions that are activated at the time

a judgment is made. A person then weighs these propositions against each other, assigns "truth values" to them (i.e., decides if these propositions should be considered valid), and tries to create consistency among them, before deciding on an explicit judgment.

A variety of factors in the introspection procedures in our studies may have contributed to a higher weighting of implicit associations for an explicit judgment following completion of the IAT. First, one could argue that we explicitly told participants that their implicit reactions were valid attitudes by asking them to report those separately. Second, participants' implicit attitudes are particularly salient after predicting IAT scores, and the particular salience of some propositions over others should raise their chances to contribute to an explicit judgment (Tesser, 1978; Tourangeau & Rasinksi, 1988). Lastly, anecdotal reading of participants' essays in Studies 3, 4 and 5, showed that after careful reflection on the differences between implicit and explicit attitudes many participants considered their implicit attitudes to be more truthful than their initial explicit thermometer ratings. In fact, we administered the same questions in Study 3 that were used as manipulation checks in Studies 1 and 2. Results showed that participants significantly agreed with the idea that their implicit attitudes reflect their "true underlying attitudes" – even though they were not told so during Study 3 (M=4.60, SD=.10), difference from the neutral mid-point of 4: $t(118) = 5.84$, $p<.001$.

Our study cannot disentangle if the completion of an implicit attitude measure such as the IAT alone would create such adaptation, or whether it was the encouragement to introspect specifically that made participants adapt their explicit attitudes to their implicit attitudes. However, research on order effects so far provides no indication that completing an implicit attitudes measure alone would cause adaptation of explicit attitudes (e.g., Nosek 2005; Hofmann,

2008). It thus seems to be *introspection* of implicit attitudes specifically that causes participants to adapt their explicit attitudes to their implicit attitudes.

This finding raises a variety of questions. That is, in light of the extensively reported bias most Americans hold against racial and ethnic minorities, and the bias most people hold in favor of their ingroup in general, this effect could be considered troublesome to the degree that participant might use their newly formed explicit attitudes to guide their behavior. On the other hand, awareness of one's own biases is considered a necessary precondition in many models of prejudice reduction (e.g., Monteith & Mark, 2005), and these models suggest that introspection of one's implicit biases is a good and healthy first step for the effortful control of prejudiced reactions. That is, participants might use their newly acquired knowledge to be more careful in their behavior, and more aware of their possibly biased reactions. During a debriefing procedure one participant pointedly noted her conflict about the newly acquired information: "I feel guilty because I think that I am an intuitive person. Yet, based on this test, it shows that if I go with my initial gut instinct about race and value judgments I am actually quite judgmental."

Our studies currently do not reveal how a person who has made the experience of holding negative associations with certain social groups is going to use that information. Our only empirical finding is an adaptation of explicit attitudes to these implicit associations. In light of these findings, it appears that it remains important to emphasize egalitarian norms. That is, participants' insights gained from introspecting implicit associations can only be useful if they are discouraged from "honestly" using those association to guide their behavior, but instead encouraged to counteract it. Future research is clearly needed to accurately assess the effects of introspection and knowledge of implicit processes on subsequent behavior, and the conditions that influence different reactions.

Relatedly, our research was only concerned with people's awareness of the content of their implicit attitudes. However, there are at least two other aspects of awareness in the realm of implicit attitudes: awareness of the source of the attitude and awareness of the impact of the attitude on subsequent behavior (Gawronski et al., 2006). Our research cannot answer questions concerning these latter kinds. However, since content awareness is a necessary precondition for both source awareness and impact awareness (it wouldn't be possible to be aware of the impact or source of an attitude of which one is not aware) the current research could be a first step into investigating the possibilities of people's awareness of their biased behaviors as well.

In the current study, we only measured implicit attitudes with the IAT– the most widely used measure of implicit attitudes. Study 4 showed that an understanding of the specific task used is not necessary to make an accurate prediction of an implicit attitude. In light of these findings it would be interesting for future research to see whether a person's prediction of their implicit attitudes maps onto other implicit attitude measures equally well.

Implicit attitude measures, and especially the IAT, have recently received considerable attention not just in social psychology (Dateline NBC, 2007; Gladwell, 2005; Operah.com, 2006; The Economist, 2012; Tierney, 2008a, 2008b). In most academic and popular representations the IAT is portrayed as measuring attitudes that are "unconscious" and inaccessibly to introspection. Both researchers and lay people alike (see Banaji, 2001; Dateline NBC, 2007; Galdwell, 2005; Tierney, 2008b) gush about the "humbling" experiences (Banaji, 2001) of discovering one's implicit attitudes. The current set of studies showed that, contrary to this widespread presentation, introspection and awareness of implicit attitudes is possible, even when people report largely different explicit "feelings" towards the same targets, and possibly even if these associations go against a person's belief system. In light of these findings, it appears

important to re-visit the presentation of our findings both to the academic community and especially to the general public.

References

Asendorpf, J.B., Banse, R., & Mücke, D. (2002). Double dissociation between implicit and explicit personality self-concept: The case of shy behavior. *Journal of Personality and Social Psychology, 83,* 380–393.

Banaji**,** M.R. (2001). Implicit attitudes can be measured. In H. L. Roediger, J. S. Nairne, I. Neath, & A. Surprenant (Eds.), The nature of remembering: Essays in remembering Robert G. Crowder (pp. 117–150). Washington, DC: American Psychological Association.

Banse, R., Seise, J., & Zerbes, N. (2001). Implicit attitudes towards homosexuality: Reliability, validity, and controllability of the IAT. *Zeitschrift fu¨r Experimentelle Psychologie, 48,* 145–160.

Bargh, J.A. (1994). The four horsemen of automaticity: Awareness, intention, efficiency, and control in social cognition. In R. S. Wyer, Jr., & T. K. Srull (Eds.), Handbook of social cognition (2nd ed.) (pp. 1-40). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Baumeister, R., & Bushman, B. (2008). *Social Psychology & Human Nature.* Belmont, CA: Thompson Wadsworth.

Bosson, J.K., Swann, W.B., & Pennebaker, J.W. (2000). Stalking the perfect measure of implicit self-esteem: The blind men and the elephant revisited? Journal of Personality and Social Psychology, 79, 631–643.

Blair, I. (2001). Implicit stereotypes and prejudice. In G. B. Moskowitz (Ed.), *Cognitive social psychology.* Mahwah, NJ: Lawrence Erlbaum Associates Inc.

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

Dateline NBC (producer)(2007, April 16). Dateline NBC: Psychological dispositions in Black & White [video webcast][Television series episode]. Retrieved May 2, 2012 from http://www.youtube.com/watch?v=sYQVDik69Nw

Devos, T (2008). Implicit attitudes 101: Theoretical and empirical insights. In WD Crano & R Prislin, Attitudes and attitude change. (p. 61-84). New York, NY: Psychology Press.

Divine, P.G. (1989). Stereotypes and prejudice: Their automatic and controlled components. *Journal of Personality and Social Psychology, 56,* 5-18.

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.

Egloff, B., & Schmukle, S.C.(2002). Predictive validity of an Implicit Association Test for assessing anxiety. *Journal of Personality and Social Psychology, 83,* 1441–1455.

Fazio, R.H., Jackson, J.R., Dunton, B.C., Williams, C.J. (1995). Variability in automatic activation as an unobtrusive measure of racial attitudes: a bona-fide pipeline? *Journal of Personality and Social Psychology, 69,* 1013-1027.

Friese, M., Hofmann, W., Schmitt, M. (2008). When and why do implicit measures predict behavior? Empirical evidence for the moderating role of opportunity, motivation, and process reliance. *European Review of Social Psychology, 19*, 285-338

Gawronski, B. (2002). What does the Implicit Association Test measure? A test of the convergent and discriminant validity of prejudice related. IATs. *Experimental Psychology, 49,* 171–180.

Gawronski, B., & Bodenhausen, G.V. (2006). Associative and propositional processes in evaluation: An integrative review of implicit and explicit attitude change. *Psychological Bulletin, 132,* 692-731.

Gawronski, B., Brochu, P. M., Sritharan, R., & Strack, F. (2012). Cognitive consistency in prejudice-related belief systems: Integrating old-fashioned, modern, aversive and implicit forms of prejudice. In B. Gawronski & F. Strack (Eds.), *Cognitive consistency: A fundamental principle in social cognition* (pp. 369-389). New York: Guilford Press.

Gawronski, B., Hofmann, W., & Wilbur, C.J. (2006). Are "implicit" attitudes unconscious? *Consciousness and Cognition, 15,* 485-499.

Gawronski, B., & LeBel, E.P. (2008). Understanding patterns of attitude change: When implicit measures show change, but explicit measures do not. *Journal of Experimental Social Psychology, 44,* 1355-1361.

Gladwell, M. (2005). Blink: The power of thinking without thinking. New York City, NY: Little, Brown and Company.

Greenwald, A.G., & Banaji, M.R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review, 102,* 4–27.

Greenwald, A.G., McGhee, D.E., Schartz, J.L.K. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology, 74,* 1464-1480.

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.

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

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

Hofmann, W., Gschwendner, T., Nosek, B. A., & Schmitt, M. (2005b). What moderates implicit-explicit consistency? *European Review of Social Psychology, 16,* 335-390.

Hofmann, W., Gschwendner, T., Wiers, R., Friese, M., & Schmitt, M. (2008). Working memory capacity and self-regulation: Towards an individual differences perspective on behavior determination by automatic versus controlled processes. *Journal of Personality and Social Psychology, 95,* 962-977.

Jones, E.E., & Sigall, H. (1971). The bogus pipeline: A new paradigm for measuring affect and attitude. *Psychological Bulletin, 76,* 349-364.

Jost, J.T., Pelham, B.W., & Carvallo, M.R. (2002). Non-conscious forms of system justification: Implicit and behavioral preferences for higher status groups. *Journal of Experimental Social Psychology, 38,* 586–602.

Kassin, S., Fein, S., Markus, H.R. (2010). Social Psychology (8th ed.). Belmont, CA: Wadsworth Cengage Learning.

Kenrick, D.T., Neuberg, S.L., & Cialdini, R.B. (2010). *Social psychology: Goals in interaction* (5th ed.). Boston, MA: Allyn & Bacon.

Kihlstron, J.F. (2004). Implicit methods in social psychology. In C Sansone, CC Morf, & AT Panter (Eds), The Sage handbook of methods in social psychology (pp. 195-212). Thousand Oaks, CA: Sage Publications.

Kim, D.-Y. (2003). Voluntary controllability of the Implicit Association Test (IAT). *Social Psychology Quarterly, 66,* 83–96.

Krueger, J. (1998). On the perception of social consensus. Advances in Experimental Social Psychology, 30, 163-240.

McConnell, A.R., Dunn, E.W., Austin, S.N., & Rawn, C.D. (2011). Blind spots in the search for happiness: Implicit attitudes and nonverbal leakage predict affective forecasting errors. Journal of Experimental Social Psychology, 47(3), 628-634.

Minear, M. & Park, D.C.(2004). A lifespan database of adult facial stimuli. *Behavior Research Methods, Instruments, & Computers. 36,* 630-633.

Monteith, M. J., & Mark, A.Y. (2005). Changing one's prejudice ways: Awareness, affect, and self-regulation. *European Review of Social Psychology, 16,* 113-154.

Nosek, B.A. (2005). Moderators of the relationship between implicit and explicit evaluation. *Journal of Experimental Psychology: General,134,* 565–584.

Nosek, B.A. (2007). Implicit-explicit relationships. *Current Directions in Psychological Sciences, 16,* 65-69

Nosek, B.A., & Banaji, M.R. (2001). The go/no-go association task. *Social Cognition, 19(6)*, 625-666.

Nosek, B.A., Banaji, M.R., & Greenwald, A. G. (2006). Website: http://implicit.harvard.edu/.

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.

Nosek, B.A., & Smyth, F.L. (2007). A multitrait-multimethod validation of the implicit association test. Implicit and explicit attitudes are related by distinct constructs. *Experimental Psychology, 54,* 15-29.

Oprah.com (producer) (2006, Jan 1). Overcoming Prejudice. *Oprah.com*. Retrieved May 2, 2012 from http://www.oprah.com/oprahshow/Overcoming-Prejudice/13

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

Quillian, L (2008). Does unconscious racism exist? Social Psychology Quarterly, 71(1), 6-11.

Ranganath (Ratliff), K.A., Smith, C.T., & Nosek, B.A. (2008). Distinguishing automatic and controlled components of attitudes from direct and indirect measurement methods. *Journal of Experimental Social Psychology*, *44*, 386-396.

Ross, L., Greene, D., & House, P. (1977). The "false-consensus effect:" An egocentric bias in social perception and attribution processes. *Journal of Experimental Social Psychology, 13,* 279-301.

Rudman, L. A., Greenwald, A.G., Mellott, D.S., & Schwartz, J.L.K. (1999). Measuring the automatic components of prejudice: Flexibility and generality of the Implicit Association Test. Social Cognition, 17, 437–465.

Rydell, R.J., & McConnell, A.R. (2006). Understanding implicit and explicit attitude change: A systems of reasoning analysis. *Journal of Personality and Social Psychology, 91,* 995-1008

Smith, E.R., & DeCoster, J. (2000). Dual-process models in social and cognitive psychology: Conceptual integration and links to underlying memory systems. *Personality and Social Psychology Review, 4,* 108-131

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

Smith, C.T., Nosek, B.A. (2011). Affective focus increases the concordance between implicit and explicit attitudes. *Social Psychology, 42,* 300-313.

Spalding, L. R., & Hardin, C. D. (1999). Unconscious unease and self-handicapping: Behavioral consequences of individual differences in implicit and explicit self-esteem. Psychological Science, 10, 535–539

Tesser, A. (1978). Self-generated attitude change. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 11, pp. 289–338): Academic.

The Economist (Author: J.F., 2012, Feb 22). You may not think what you think you think. *The Economist.* Retrieved May 2, 2012 from http://www.economist.com/node/21548123

Tierney, J. (2008a, Nov. 17). In bias test, shades of gray. *The New York Times.* Retrieved May 2, 2012, from http://www.nytimes.com/2008/11/18/science/18tier.html

Tierney, J. (2008b, Nov. 18). A shocking test of bias. The New York Times. Retrieved May 2nd, 2012 from http://tierneylab.blogs.nytimes.com/2008/11/18/a-shocking-test-of-bias/

Tourangeau, R., & Rasinski, K.A. (1988). Cognitive processes underlying context effects in attitude measurement. *Psychological Bulletin, 103*, 299–314.

Uhlmann, E.L., & Nosek, B.A. (in press). My culture made me do it: Lay theories of responsibility for automatic prejudice. Social Psychology.

Uhlmann, E.L., Poehlman, T.A., & Nosek, B.A. (in press). Automatic associations: Personal attitudes or cultural knowledge?. In J. Hanson (Ed.), Ideology, Psychology, and Law. Oxford, UK: Oxford University Press.

Wilson, T.D., Dunn, D.S., Bybee, J.A., Hyman, D.B., Rotondo, J.A. (1984). Effects of analyzing reasons on attitude-behavior consistency. *Journal of Personality and Social Psychology, 47,* 5-16

Wilson, T.D., Dunn, D.S., Kraft, D. & Lisle, D.J. (1989). Attitude change, and attitude-behavior consistency: The disruptive effects of explaining why we feel the way we do. In: L. Berkowitz (Ed.), Advances in experimental social psychology (Vol. 22, pp. 287-343). Orlando, FL: Academic Press.

Wilson, T.D., Lindsey, S., Schooler, T.Y. (2000). A model of dual attitudes. *Psychological Review, 107,* 101-126

Appendices

Appendix A: Valence words used in all IATs

Positive words	Negative words
PLEASANT	HORRIBLE
DELIGHT	ANGRY
HELPFUL	TERRIBLE
JOY	TRAGIC
BEAUTIFUL	HATE
SMILE	DESTROY
WONDERFUL	BRUTAL
ENJOY	EVIL
CHEERFUL	DISASTER
SUCCESS	UGLY

Appendix B 1: Writing Task (IAT training) "True Attitudes Condition" Studies 1 and 2:

Do You Know Yourself?

The goal of the next part of the study is to measure how well you know yourself, specifically what you think is good and bad. This may seem like an odd question because who knows you better than yourself? You know what foods you like and dislike, which of your acquaintances are friends and which are not, etc. However, there are many things about which you may not be so sure of. For example, you may not be sure if you like a place that you haven't been to yet; you may have conflicting feelings about something (e.g., ice-cream tastes great but I think it makes me fat); or you may really want to be one way even though you feel differently (I don't really feel comfortable around dogs, but as an animal lover I *know* that I like dogs).

These examples suggest that a person's true attitude can sometimes be difficult to figure out, even for the person. One of the ways psychologists have solved this problem is to distinguish between **a person's explicit attitude** (what they think of themselves) and **the person's implicit attitude** (the true, underlying attitude that might not be consciously known). In some cases a person's explicit and implicit attitudes are the same, but in other cases they can be quite different. One way to ask yourself if you know your implicit attitudes is to pay attention to your initial gut feeling towards something before you have time to think or reason about your beliefs. The goal of this part of the study is to see how well you know – or can guess – what your implicit attitudes are toward different social groups, regardless of what you would like to think about them.

To make sure you understand this part, in your own words, please briefly describe the difference between **implicit and explicit attitudes** as you understand it.

Original text box is bigger

Measuring Implicit Attitudes

We have found that people do better at this task once they understand how we measure implicit attitudes. How would you measure something that you might not be conscious of?

The test we will use in this study is called **the Implicit Attitude Test (IAT)** and it reveals a person's attitude by how easy or difficult it is for them to complete certain tasks. As an example, imagine that there is one deck of cards that has a picture of either a flower or an insect printed on each one, and a second deck of cards that has either a pleasant word (e.g., BEAUTIFUL) or an unpleasant word (e.g., DEATH) printed on each one. Imagine further that we shuffle these two decks of cards together and then ask you to sort them into two piles, as fast as you can. Would it be easier for you to put the flower pictures with the good words and the insect pictures with the bad words, or would it be easier for you to put the insect pictures with the good words and the flower pictures with the bad words? For a person who likes flowers more than insects, the "flower+good/ insect+bad" sorting task is easier than the alternative, "insect+good/ flower+bad". This is also typically one of those cases in which a person's implicit attitude is pretty similar to his/her explicit attitude (i.e., people who explicitly like flowers, also tend to implicitly like them).

Although you won't actually sort cards in this study, we will measure your implicit attitudes with a computer test that operates in much the same way.

Before you try this, however, please try to phrase in your own words your understanding of the **Implicit Attitude Test (IAT).** How does it work? How does it measure a person's implicit attitude?

Once you have completed this writing task, please return to your computer screen and press the "next" button.
Original text box is bigger

Appendix B 2: Writing Task (IAT training) "Culturally Learned Associations" condition (Studies 1 and 2):

Do You Know How Your Culture Influences You?

The goal of the next part of the study is to see how well you know your culture and its influence on you, specifically what your culture tells you is good and bad. Most Americans, for example, accept their culture's values that being independent is good and lying is bad. Once accepted, these values become part of a person's **explicit attitudes** (i.e., what the person truly believes and tries to achieve).

However, there may be parts of a culture that at least some people do not accept. For example, the size and shape of fashion models and professional athletes suggest how people "ought" to look even though it is a cultural ideal that is nearly impossible to meet, and this can contribute to eating disorders and depression. Each one of us makes decisions about which parts of our culture we accept and which parts of our culture we do not accept.

What is interesting for psychologists is that even though people may not agree with something in their culture, it can still "get inside" their head. For example, you may consciously know that a fashion model's body is not a healthy or realistic ideal, and yet on some level you may worry about not looking as "good." This more subtle influence of culture is something that is referred to as an **implicit cultural association** (i.e., something from the culture that affects you on some level though you may not be totally aware of it).

In some cases a person's explicit attitude will be the same as their implicit cultural association, but in other cases they can be quite different. One way to ask yourself if you know your implicit cultural associations is to pay attention to your initial gut feeling towards something before you have time to think or reason about your beliefs. The goal of this part of the study is to see how well you know – or can guess – what implicit cultural associations you have with regard to different social groups, regardless of what you truly believe about them explicitly.

To make sure you understand this part, in your own words, please briefly describe the difference between **implicit cultural associations** and **explicit attitudes** the way you understand it.

Original Text box is bigger

Measuring Implicit Cultural Associations

We have found that people do better at this task once they understand how we measure implicit cultural associations. How would you measure something that you might not be conscious of?

The test we will use in this study is called the **Implicit Association Test (IAT)** and it reveals associations by how easy or difficult it is for people to complete certain tasks. As an example, imagine that there is one deck of cards that has a picture of either a flower or an insect printed on each one, and a second deck of cards that has either a pleasant word (e.g., BEAUTIFUL) or an unpleasant word (e.g., DEATH) printed on each one. Imagine further that we shuffle these two decks of cards together and then ask you to sort them into two piles. Would it be easier for you to put the flower pictures with the good words and the insect pictures with the bad words, or would it be easier for you to put the insect pictures with the good words and the flower pictures with the bad words? Because our culture is more favorable toward flowers than insects (have you ever sent a loved one an insect for Valentine's Day or Mother's Day?!), the "flower+good/ insect+bad" sorting task is usually easier than the alternative, "insect+good/ flower+bad". This is also typically one of those cases in which the implicit cultural association is pretty similar to people's explicit attitude (i.e., most people accept the cultural idea that flowers are better than insects).

Although you won't actually sort cards in this study, we will measure implicit associations with a computer test that operates in much the same way.

Before you try this, however, please try to phrase in your own words your understanding of the **Implicit Associations Test (IAT).** How does it work? How does it measure a person's implicit cultural association?

Once you have completed this writing task, please return to your computer screen and press the "next" button.

Original text box is bigger

Appendix B 3: Additional task in Studies 2, 3, 5, and the full-explanation conditions in Study 4. Task was completed after participants predicted, completed, and received feedback on an insect-flower IAT and a dog-cat IAT). It is supposed to encourage reflection on the meaning of implicit attitudes:

Writing Exercise #2

Now that you have predicted your score, completed, and received feedback on two IATs, we would like you to step back and take a second to think about what your results mean.

What does it mean to you to say that you have a certain true implicit attitude [a certain culturally learned association] that is _____ *(your score)* when comparing FLOWERS to INSECTS, and _____ *(your score)* when comparing CATS to DOGS?

In your own words, please explain what your results mean.

Once you have completed this writing exercise, press "6" on your keyboard to continue with the experiment. Press "6" only when you have completed this writing exercise.

Original text box is bigger

Appendix B 4: Writing Task (IAT training) in Studies 3, 5, and the full explanation conditions in Study 4

Do You Know Yourself?

The goal of the next part of the study is to measure how well you know yourself, specifically what you think is good and bad. This may seem like an odd question because who knows you better than yourself? You know what foods you like and dislike, which of your acquaintances are friends and which are not, etc. However, there are many things about which you may not be so sure of. For example, you may not be sure if you like a place that you haven't been to yet; you may have conflicting feelings about something (e.g., ice-cream tastes great but I think it makes me fat); or you may really want to be one way even though you feel differently (I don't really feel comfortable around dogs, but as an animal lover I *know* that I like dogs).

These examples suggest that a person's attitude can sometimes be difficult to figure out, even for the person. One of the ways psychologists have solved this problem is to distinguish between **a person's explicit attitude** (what you think you like once you've had time to think and reflect about it) and **the person's implicit attitude** (the underlying attitude that gets triggered spontaneously and that might not be consciously known). In some cases a person's explicit and implicit attitudes are the same, but in other cases they can be quite different. One way to ask yourself if you know your implicit attitudes is to pay attention to your initial gut feeling towards something before you have time to think or reason about your beliefs. The goal of this part of the study is to see how well you know – or can guess – what your implicit attitudes are toward different social groups, regardless of your explicit attitudes. That is, can you tell what spontaneous attitude you hold independent of what you would say once you had time to think and reflect about it?

To make sure you understand this part, in your own words, please briefly describe the difference between **implicit and explicit attitudes,** as you understand it.

Original text box is bigger

Measuring Implicit Attitudes

We have found that people do better at knowing their implicit attitudes once they understand how we measure them. How would you measure something that you might not be conscious of?

The test we will use in this study is called **the Implicit Associations Test (IAT)** and it reveals a person's implicit attitude by how easy or difficult it is for them to complete certain tasks. As an example, imagine that there is one deck of cards that has a picture of either a flower or an insect printed on each one, and a second deck of cards that has either a pleasant word (e.g., BEAUTI-FUL) or an unpleasant word (e.g., DEATH) printed on each one. Imagine further that we shuffle these two decks of cards together and then ask you to sort them into two piles, as fast as you can. Would it be easier for you to put the flower pictures with the good words and the insect pictures with the bad words, or would it be easier for you to put the insect pictures with the good words and the flower pictures with the bad words? For a person who implicitly likes flowers more than insects, the "flower+good/ insect+bad" sorting task is easier than the alternative, "insect+good/ flower+bad". This is also typically one of those cases in which a person's implicit attitude is pretty similar to his/her explicit attitude (i.e., people who explicitly like flowers when they think about it, also tend to implicitly like them spontaneously).

Although you won't actually sort cards in this study, we will measure your implicit attitudes with a computer test that operates in much the same way.

Before you try this, however, please try to phrase in your own words your understanding of the **Implicit Associations Test (IAT).** How does it work? How does it measure a person's implicit attitude?

Once you have completed this writing task, please call the experimenter so you can continue the study.

Original text box is bigger

Appendix C: Means for predictions, Thermometer Ratings, and IAT scores

Study	Condition	target pair	Prediction 1-7	IAT *D* scores	Therm. Ratings pre	Therm. Ratings post
Study 1	"Associations"	Asian-White	4.85	0.36	65	63
		Black-White	5.03	0.39	74	69
		Celebrity-Regular	3.47	-0.12	52	65
		Child-Adult	3.09	-0.17	77	78
		Latino-White	5.06	0.37	65	64
	"Attitudes"	Asian-White	5.06	0.35	72	73
		Black-White	5.23	0.41	80	78
		Celebrity-Regular	4.16	-0.11	56	64
		Child-Adult	3.23	-0.22	82	83
		Latino-White	5.35	0.48	72	69
Study 2	"Associations"	Asian-White	4.98	0.36	70	67
		Black-White	4.96	0.34	76	73
		Celebrity-Regular	4.04	-0.08	56	62
		Child-Adult	3.36	-0.11	84	82
		Latino-White	5.27	0.32	66	65
	"Attitudes"	Asian-White	4.98	0.29	72	71
		Black-White	4.89	0.37	80	78
		Celebrity-Regular	4.02	0.03	60	68
		Child-Adult	3.56	-0.23	83	82
		Latino-White	5.31	0.34	71	70
Study 3	Self rating first	Asian-White	4.83	0.38	15	16
		Black-White	4.82	0.36	6	12
		Celebrity-Regular	3.58	-0.14	22	6
		Child-Adult	3.23	-0.21	-4	-8
		Latino-White	4.92	0.40	15	18
	Other Rating first	Asian-White	4.70	0.30	13	18
		Black-White	4.91	0.34	7	15
		Celebrity-Regular	3.13	-0.17	14	-1
		Child-Adult	2.96	-0.21	-4	-9
		Latino-White	5.01	0.39	15	21

Thermometer ratings are computed as absolute (i.e., non-relative) ratings of the first target group in a pair for Studies 1 and 2; and higher ratings mean more positive evaluations of the first target group (Asian, Black, Celebrity, Child, or Latino). For Studies 3-5, thermometer ratings are computed as relative difference scores of ratings of the second target group minus ratings of the first target group; higher ratings indicate a preference for the second target group (White, celebrity, or adult) over the first.

Study	Condition	target pair	Prediction	IAT scores	Therm. Ratings pre	Therm. Ratings post
Study 4	Full explanation/ full experience	Asian-White	4.73	0.32	13	18
		Black-White	4.59	0.42	6	11
		Celebrity-Regular	3.58	-0.11	15	1
		Child-Adult	3.24	-0.23	-7	-14
		Latino-White	4.98	0.42	14	17
	Full explanation/ no experience	Asian-White	4.72	0.27	9	12
		Black-White	4.77	0.42	6	8
		Celebrity-Regular	3.77	-0.14	19	6
		Child-Adult	3.41	-0.27	-3	-5
		Latino-White	4.91	0.35	16	17
	Minimal explanation/ full experience	Asian-White	4.57	0.33	13	13
		Black-White	4.74	0.35	9	12
		Celebrity-Regular	4.04	-0.11	20	6
		Child-Adult	3.25	-0.21	-6	-8
		Latino-White	4.88	0.39	17	17
	Minimal Explanation/ no experience	Asian-White	4.7	0.39	12	14
		Black-White	4.76	0.43	8	12
		Celebrity-Regular	3.82	-0.15	17	6
		Child-Adult	3.19	-0.27	-8	-10
		Latino-White	4.9	0.44	15	15
Study 5	Reasons	Asian-White	4.78	0.28	12	16
		Black-White	4.83	0.49	7	11
		Celebrity-Regular	3.92	-0.08	23	13
		Child-Adult	3.07	-0.29	-8	-12
		Latino-White	4.81	0.3	8	16
	Control	Asian-White	4.93	0.4	15	23
		Black-White	4.86	0.4	10	20
		Celebrity-Regular	3.77	-0.15	28	10
		Child-Adult	3.49	-0.26	-6	-12
		Latino-White	5.18	0.38	16	25
	Intuitive	Asian-White	4.43	0.35	9	12
		Black-White	4.63	0.36	7	10
		Celebrity-Regular	3.62	-0.24	18	3
		Child-Adult	3.31	-0.18	-6	-6
		Latino-White	4.74	0.33	14	16

For Studies 3-5, thermometer ratings are computed as relative difference scores of ratings of the second target group minus ratings of the first target group; higher ratings indicate a preference for the second target group (White, celebrity, or adult).