European Journal of Social Psychology, *Eur. J. Soc. Psychol.* **42**, 189–199 (2012)
Published online 18 November 2011 in Wiley Online Library (wileyonlinelibrary.com) **DOI**: 10.1002/ejsp.859

Research article

The IAT is sensitive to the perceived accuracy of newly learned associations

ERIC SIEGEL[1], HAROLD SIGALL[1] AND DAVID E. HUBER[2]*
[1]*Department of Psychology, University of Maryland, College Park, USA;* [2]*Department of Psychology, University of California, San Diego, La Jolla, USA*

Abstract

Three experiments tested whether the Implicit Association Test (IAT) is sensitive to the perceived accuracy of newly learned associations. In experiment 1, participants learned to associate positive or negative attributes with two novel groups. Participants in one condition were told that the attributes accurately described the groups; in a second condition, prior to learning, they were made aware that the attributes were randomly assigned to the groups. Participants were given an IAT and an explicit measure testing attitudes towards the two groups. When the participants were told that the attributes were accurate, their IAT performance and explicit measure responses indicated a preference for the more positively described group but when the attributes were known to be arbitrary, preferences were reduced according to both measures. Experiment 2 replicated these results and demonstrated that the associations were learned even in the random condition. Experiment 3 included a condition that placed "not" before each attribute, which demonstrated that people can incorporate a negative modifier into a learned association. Explicit attitudes and the IAT showed reversed preferences in this negation condition. These experiments imply that the IAT is sensitive to the perceived accuracy of learned associations. Copyright © 2011 John Wiley & Sons, Ltd.

The study of attitudes is one of the central concerns of social psychology. An *attitude* is defined as an evaluative expression of an object or person. Attitudes have traditionally been thought of as consciously recognized constructs characterized by beliefs, behavioral intentions, and evaluations of attitude objects (Eagly & Chaiken, 1998). In recent years, researchers have developed dual attitude models. These models suggest that there is a sharp separation between the conscious and unconscious components of attitude. According to these theories, there are two distinct types of attitudes: *Implicit attitudes*, which exist outside of awareness, are activated automatically, require conscious effort to suppress, and are difficult to change. In contrast, *explicit attitudes* are constructed on the spot using whatever relevant information is consciously available and therefore require psychological effort to be activated and maintained (e.g., Wilson, Lindsey, & Schooler, 2000).

Implicit attitude theories have led to the development of implicit measures. These measures are useful assessment tools because people are less able to hide their attitudes on an implicit measurement task. One of the most widely used implicit measures is the Implicit Association Test (IAT) introduced by Greenwald, McGhee, and Schwartz (1998). Demonstrating the popularity of the IAT, a recent meta-analysis by Greenwald, Poehlman, Uhlmann, and Banaji (2009) assessed the predictive validity of the IAT across 122 research reports. The IAT compares attitudes towards two groups by measuring the association between the groups and positive and negative evaluations. When taking the IAT, participants must categorize four groups of items: positive adjectives, negative adjectives, and two distinct groups (e.g., Black faces and White faces). The IAT measures the attitude preferences towards the two groups by comparing reaction times between two blocks of trials. In one block, called the compatible block, the response to the preferred group uses the same response key as responses to positive adjectives. In the incompatible block, the response to the preferred group uses the same response key as responses to negative adjectives. People tend to respond more slowly in the incompatible block compared with the compatible block. This slowdown is called the IAT effect, and it can be measured by subtracting the average reaction time of the compatible block from the average reaction time of the incompatible block. This difference is used to determine the relative evaluation of the two groups (i.e., which group is preferred or, alternatively, disliked the least), although researchers have developed more advanced scoring algorithms designed to filter out extraneous factors that might influence a simple difference score (e.g., Greenwald, Nosek, & Banaji, 2003).

It has been proposed that the IAT effect results from greater response conflict in the incompatible block compared with the compatible block, resulting in slower reaction times during the incompatible block to avoid errors (Greenwald et al., 2003). Response conflict arises from the associations held by the participants. In the incompatible block, the favored group

*Correspondence to: David E. Huber at the Department of Psychology, University of California, San Diego, 9500 Gillman Dr., La Jolla, CA 92093–0109, USA.
E-mail: dhuber@ucsd.edu

Received 15 December 2010, Accepted 24 May 2011

shares a response key with the negative adjectives, and thus a stimulus from the favored group lends itself to two competing responses: one reflecting the group membership response and the other reflecting the association between that group and a positive evaluation response. In contrast, for the compatible block, the two responses elicited by a stimulus from the favored group indicate the same response key, and so there is no response conflict. A number of alternative explanations have been proposed for the process underlying the IAT effect. For instance, researchers have found that IAT performance is affected by the salience of the tested groups, indicating that the IAT effect partially reflects different degrees of familiarity for each group rather than associations between the groups and valence (Brendl, Markman, & Messner, 2001; Rothermund & Wentura, 2004). Other studies have found evidence that the IAT is sensitive to the associations between groups and cultural, rather than personal, constructs (Fazio, Han, & Olson, 2006; Olson & Fazio, 2004). Beyond these studies that question the type of association underlying the IAT, other studies have found that the magnitude of an individual's IAT score is partially determined by individual differences in cognitive ability (e.g., Blanton & Jaccard, 2006; Klauer, Schmitz, Teige-Mocigemba, & Voss, 2010).

Although these alternative explanations are disputed (e.g., Greenwald, Nosek, & Sriram, 2006), proponents on both sides of the debate agree that the direction of the IAT effect (i.e., whether the difference score is positive or negative, indicating a preference for one group or the other) is due to differences in the latent associations attached to each group, regardless of whether these associations are cultural, personal, or a sense of familiarity. Critically, it is now understood that these associations must be activated during the IAT to produce response conflict, and it has been found that mindset manipulations can change the IAT effect (Han, Czellar, Olson, & Fazio, 2010). Putting aside the manner in which these associations relate to attitudes, we seek to gain a better understanding of the associations that underlie an IAT effect. Because the IAT is typically used to assess previously learned associations (i.e., participants enter the lab with a lifetime of experience regarding the groups tested in the IAT), little is known about the types of learning that do or do not create the associations necessary to produce response conflict while taking the IAT. In the current study, we ask whether newly learned associations give rise to an IAT effect even if those associations are known to be inaccurate or arbitrary. In other words, we ask whether the IAT is sensitive to the perceived accuracy of newly learned associations.

Karpinski and Hilton (2001) performed one of the earliest studies examining the influence of newly learned associations on the IAT. They argued that the IAT reflects the information people are exposed to rather than how they feel about that information:

> According to the environmental association model of the IAT, a high score on a White/Black IAT, for example, should not be seen as indicating that the individual has more favorable evaluations of Whites compared with Blacks. Instead, the score may simply indicate that the individual has been exposed to a larger number of positive-White and negative-Black associations than negative-White and positive-Black associations (Karpinski & Hilton, 2001, p.776).

To test this assertion, Karpinski and Hilton had participants learn word pairs for a later memory test. These word pairs consistently paired the elderly with positive adjectives and young people with negative adjectives, and the instructions were simply to memorize the word pairs. After this learning procedure, Karpinski and Hilton gave participants an elderly/youth IAT and found that the IAT was influenced by the word pairs, reversing the negative bias towards the elderly normally found by the IAT. These findings demonstrated that information can influence IAT results merely through exposure to associations. This suggests that the IAT reflects learned associations regardless of the respondents' beliefs about those associations. However, because Karpinski and Hilton used pre-existing groups, it is likely that the participants had both positive and negative associations about the groups prior to entering the lab, in which case the effect of the word pairs may have been to prime (i.e., activate) previously learned associations rather than to be a direct implantation of new associations. To prevent latent associations from influencing the results, it is necessary to use novel groups and create new positive or negative associations with those groups.

To address these issues without contamination from latent associations, Gregg, Seibt, and Banaji (2006) examined whether the IAT effect depends on the test taker's beliefs regarding novel fictional groups. They noted conflicting evidence within the literature; some studies found that the IAT was immune to respondents' beliefs (e.g., Banse, Seise, & Zerbes, 2001; Gawronski & Strack, 2004), whereas others found that IAT effects are changed by new information (e.g., Blair, Ma, & Lenton, 2001; Dasgupta & Greenwald, 2001). Gregg et al. addressed this inconsistency by exploring the processes giving rise to IAT effects and change in IAT effects. Participants learned the attributes of novel fictional groups in a concrete manner (e.g., adjective–name word pairs) or abstract manner (by supposition). Regardless of the method, these newly learned associations produced an IAT effect. In their third experiment, they repeated the concrete learning procedure, and after completion of an initial IAT, participants were told that a computer error had reversed the adjectives associated with each group during the learning task. Thus, after the fact, they were told that the associations were inaccurate. A second IAT was found to reflect the initially learned associations even though measures of explicit attitude showed a preference change in light of the "computer error." In a fourth experiment, they gave participants the concrete learning procedure and then gave them an entirely new version of that procedure designed to reverse the initial associations. Once again, the IAT reflected the initially learned associations.

Gregg et al. (2006) argued that these results demonstrate that implicit attitudes, once formed, are resistant to change, whereas explicit attitudes can be changed in the face of new information. However, there are certain elements of their methodology that make interpretation of these results difficult. More specifically, in the experiments that produced a dissociation between the IAT and explicit measures of attitude, participants were given two competing sources of information.

This occurred because participants were first taught the associations (once source of information) and *subsequently* learned that the original associations were incorrect (a second source of information). These results suggest that the IAT is insensitive to the perceived accuracy of newly learned associations, but this conclusion has only been demonstrated when the accuracy of the associations is questioned after the fact. With two sources of information, it may simply be that there was a primacy effect (Anderson & Barrios, 1961) such that the first set of associations (e.g., the initial word pairs) had a larger influence on the IAT than the second set of associations (e.g., knowledge of the computer error or learning of reversed associations).[1] In addition to using two sources of information, their Experiment 3 used two separate IATs, and it is possible that performance on the second test was heavily affected by the first IAT (e.g., perhaps the first test served to "consolidate" IAT performance, making it resistant to other influences). In the current study, we avoided these complications by using just one IAT and just one set of associations to re-examine the issue of whether the IAT is sensitive to the perceived accuracy of newly learned associations; this was done by manipulating the accuracy of the associations *before they were learned*.

De Houwer (2006) advocated a different interpretation of the Gregg et al. (2006) results. He noted that there was an equally large IAT effect for both the concrete word-pair learning condition and the abstract supposition condition in which participants were asked to suppose that the groups had certain characteristics without actually viewing word pairs. He argued that these IAT effects reflected participants' conscious intentions (what De Houwer termed "propositional processes") rather than unconscious associations. Using the paradigm of Gregg et al., he taught participants about novel fictional groups by instructing them that these groups would be predictive (at some later time) of positive (or negative) pictures:

> In this experiment, you will see pleasant, positive photos (e.g., of flowers) and unpleasant, negative photos (e.g., of mutilated bodies). Each photo will be preceded by a fictional group that indicates which type of photo (positive or negative) will be presented. It is very important that you remember which group goes together with which kind of photo. You need this information to complete the task successfully. This information will not be presented again later on, so remember well which group goes together with which kind of photo. (De Houwer, 2006, p. 181)

Participants never saw the photos, and yet there was an IAT effect after receiving these instructions. Critically, unlike the Gregg et al. study, De Houwer did not tell participants that the photos described the associated groups but merely that they were to remember which group would be paired with positive versus negative photos. Because this pairing was simply something to memorize, this result implies that IAT effects can occur regardless of the perceived accuracy of an association. However, there are two problems with this conclusion. First, it is unclear what participants thought of these associations and whether they used them to develop attitudes towards the groups. This could have been assessed with an explicit measure of attitude. Second, because De Hower's study only included one condition, it is unclear whether the IAT effect would have been substantially larger if participants were instructed that the associations were true attributes of the fictional groups. By including two conditions, the current study assessed the effect of perceived accuracy.

The current research determined whether the IAT is sensitive to the perceived accuracy of newly learned associations. Previous attempts to answer this empirical question had methodological limitations that we believe are resolved with the current paradigm. We taught participants about the fictional groups created by Gregg et al. (2006). However, unlike their study, we manipulated the accuracy of the associations before they were learned so that there was one source of information, and we used only one IAT to avoid contamination from prior tests. As in De Houwer's (2006) experiment, one of our conditions made it clear that the associations were completely arbitrary. This was reinforced by letting participants flip their own coin to determine which group would be paired with positive versus negative adjectives. However, unlike De Houwer's study, we also included a condition in which the associations were thought to be valid, and we collected explicit measures of attitude for both conditions. Although the primary concern of this research is methodological, the results have important theoretical implications. However, these implications depend on whether one subscribes to a single or dual model of attitudes—whether the IAT is an *implicit measure* of attitudes or whether it is a measure of *implicit attitudes*. In the general discussion, we consider the results in light of these theoretical alternatives.

EXPERIMENT 1

Participants were taught to associate members of two novel groups with either positive or negative adjectives. In the random condition, participants flipped a coin to determine which set of adjectives was associated with each group. In the accurate condition, participants were told that the adjectives accurately described the groups. In both cases, participants were instructed to learn the pairings of adjectives and group members and that they would be tested on these associations.

Besides the accuracy manipulation, the strength of the associations was manipulated to test Karpinski and Hilton's (2001) assertion that the IAT reflects the degree of exposure to information. We manipulated the strength of the associations by exposing participants either to 120 or 240 trials during initial learning. The choice of 240 trials for the high exposure condition was based on the procedure of Gregg et al. (2006), which likewise used 240 trials. The low exposure condition was then set at half this number of trials. Following initial learning, we gave participants an IAT and an explicit attitude measure featuring the two novel groups.

[1]It is important to note that even if a primacy effect explains the failure to change IAT scores with subsequent information in the Gregg et al. (2006) study, this implies that the explicit ratings were not subject to the same primacy effect. Thus, the Gregg et al. results could still be viewed as supporting a dual attitude model under this alternative explanation.

Participants

Forty-five University of Maryland undergraduates, 31 women and 14 men, participated in the experiment. They received extra credit for their participation.

Procedure

Participants were seated in a small room containing a computer and given oral instructions complemented by on-screen text. They were told that the purpose of the experiment was to test how knowledge about a group affects one's ability to identify and categorize group members. The experimenter explained that the experiment used groups that the participant had never heard of before in order to prevent the influence of prior knowledge and/or bias. The two groups were called the Luupites and the Niffites.

Participants were told that they would learn about the groups through an exercise in which each group would be paired with a set of adjectives. The IAT was described as the "testing phase" of the experiment in which the participants were tested to see if they had learned which adjectives were associated with each of the groups. After receiving these instructions, the participants began the learning task. After completing the learning phase, the participants completed an IAT featuring the two groups. Lastly, they completed a short explicit measure that tested attitudes towards the two groups. Besides the general instructions, specific instructions concerning the accuracy manipulation and how to respond in the various sections of the IAT were provided and are described below.

There were two independent variables included in the study. The first was the accuracy manipulation, whether the participants were told the adjectives described the groups. The second was the strength of the learned associations, represented by the number of trials included in the learning task. In addition to the main independent variables of accuracy and associative strength, the order of the compatible and incompatible blocks in the IAT was counterbalanced, as were the specific pairings of groups and adjectives. The type of measure (IAT vs. explicit) was also treated as an independent variable. The experimental design was mixed, with type of measure as a within-subject variable and accuracy (accurate vs. random pairings), associative strength (long vs. short task), IAT block order, and the pairing of the adjective/group associations (Luupite-positive vs. Niffite-positive) as between-subject variables. Because the IAT block order and adjective group associations variables were included for methodological and not theoretical reasons, they were not included in the final analysis once it was determined that they did not have a significant effect on the results.

The Learning Task

The learning task introduced participants to the Luupites and the Niffites and formed the associations tested by the measures. We utilized the fictional groups designed by Gregg et al. (2006) to eliminate the possibility that pre-existing attitudes would influence the experimental results. In pre-tests, Gregg et al. found that initial attitudes towards these groups and their members were neutral. There were eight members of each group, with group membership easily identifiable by the structure of the names: Luupite names all contained double vowels and ended in -lup (e.g., Neenalup, Maasolup), whereas Niffite names all contained double consonants and ended in-nif (e.g., Eskannif, Lebbunif). Participants were told that in reality these were ancient historical groups but that the groups' names and the names of the group members had been changed so that they could not be recognized.

The learning task consisted of a series of random trials, similar in format to the IAT, of which half featured Luupite names and the other Niffite names. Participants classified the names on the basis of their group by pressing the appropriate key on a response box with millisecond precision. If the participants classified a name incorrectly, a red "X" appeared at the bottom of the screen, and participants could not proceed until they correctly classified the name. The names of the two groups were displayed in the upper corners of the screen, and their location (right or left corner) corresponded to the answer key assignments, which were randomly changed on each trial to prevent association of the groups with a specific answer key.

To familiarize the participants with the format of the task, they were given 20 practice trials to categorize just the names. After this practice phase, the main learning task began, during which each name was immediately preceded by an adjective which was flashed on the screen for 200 milliseconds. Throughout the entire learning phase, the adjectives from one set (positive or negative) were consistently paired with names from one group (Niffite or Luupite). Half of the trials were Niffite trials, and half were Luupite trials, in random order. There were eight positive adjectives and eight negative adjectives that preceded the paired group names. On every trial, an adjective was randomly selected, with replacement, and a name was randomly selected, with replacement. The learning task consisted of 120 trials in the *short* condition or 240 trials in the *long* condition.

In the *accurate* condition, participants were instructed, prior to learning, that the adjective sets accurately described the groups they were paired with in the learning task and that the adjectives could be used to form an accurate impression of the two groups. Because they had no knowledge to the contrary, we assumed that the participants in this condition would believe the adjectives accurately described the groups.

In the *random* condition, the participants were instructed prior to learning that even though each adjective set was paired with a specific group, they did not necessarily accurately describe the associated group. In order to convince participants that the adjective pairings were arbitrary, participants were asked to flip a coin to determine which adjective set was to be associated with which group during the learning task.

The Implicit Association Test

The participants' implicit attitudes were assessed using a Luupite/Niffite, good/bad IAT using adjectives similar to the ones featured in the learning task. The format of the IAT was identical to the seven-block procedure recommended by

Greenwald et al. (2003), featuring Luupites and Niffites as the category groups.

For 22 participants, the compatible block was presented in block 3 of the IAT, and the incompatible block was presented in block 5, whereas for 23 participants the reverse was true.

Explicit Measure of Attitudes

Using the number pad on the keyboard rather than the response box, participants rated how they felt about the two groups. They were first asked to complete the statement "I think the [Luupites/Niffites] are..." on a seven-point scale ranging from "very bad" (1) to "very good" (7). Additionally they were asked to complete "I like the [Luupites/Niffites]..." on a seven-point scale ranging from "not at all" (1) to "very much." (7). Participants gave answers to both of these questions in regard to each group separately.

RESULTS

Data Reduction

The Cronbach's alpha between the two explicit measure questions was sufficiently high ($\alpha = .82$) to combine the two measures, so the scores on the two explicit measures were averaged together.

Standardized IAT scores were calculated using the D_1 algorithm recommended by Greenwald et al. (2003). For the reported IAT scores, a positive number indicates a greater preference for the group paired with positive adjective over the group paired with negative adjectives. The D_1 transformation takes the difference in average reaction time between the mixed blocks and divides this difference by the standard deviation pooled across these blocks. Therefore, the IAT scores are on a *z*-scale with 0 representing no difference and positive or negative values representing relative differences expressed in units of standard deviation. For the explicit ratings, a separate average score was calculated for the positive and negative groups for each participant. The score for the group paired with negative adjectives was subtracted from the score for the group paired with positive adjectives. This difference was used for most analyses but for correlations between the IAT and explicit ratings, this difference was z-transformed across participants and conditions to place IAT scores and explicit rating on a common scale.

Analysis

Table 1a reports average IAT reaction times for each block, Table 1b reports the untransformed difference scores for both the IAT and explicit measures, and Table 1c reports the standardized versions of these difference scores.

Preliminary analyses showed that neither IAT block order nor adjective–group pairing had an effect on IAT performance, and these nuisance variables were ignored in subsequent analyses. Preliminary analyses of associative strength did not find any main effects or interactions that involved associative strength, and associative strength was ignored in subsequent analyses.

Table 1a. Experiment 1 IAT response times

	Block			
	Compatible		Incompatible	
Accurate condition	Mean (milliseconds)	Standard deviation (milliseconds)	Mean (milliseconds)	Standard deviation (milliseconds)
Accurate	888	316	1084	400
Random	832	233	884	236

Table 1b. Experiment 1 untransformed differences

	Measure			
	IAT		Explicit	
Accurate condition	Mean (milliseconds)	Standard deviation (milliseconds)	Mean	Standard deviation
Accurate	196	240	4.50	1.87
Random	52	199	0.31	0.99

Table 1c. Experiment 1 standardized differences

	Measure			
	IAT		Explicit	
Accurate condition	Mean	Standard deviation	Mean	Standard deviation
Accurate	0.56	0.56	0.96	0.74
Random	0.24	0.49	−0.70	0.39

As seen in Tables 1b for the explicit measure, participants demonstrated a greater preference for the positive group in the accurate condition than in the random condition [$t(43) = 9.73$, $p < .01$]. A one-sample *t*-test indicated that the average in the random condition was not significantly different from zero [$t(25) = 1.59$, $p = .12$]. In contrast, the accurate condition was significantly greater than zero [$t(18) = 10.49$, $p < .01$].

The IAT also revealed a greater preference for the positive group in the accurate condition compared with the random condition [$t(43) = 2.08$, $p < .05$]. Unlike the explicit measure, a one-sample *t*-test indicated that the average IAT effect in the random condition was significantly different from zero [$t(26) = 2.45$, $p < .05$]. The accurate condition mean was also significantly different from zero [$t(18) = 4.40$, $p < .01$].

In summary, the pattern of results was similar for the explicit measure and the IAT: both measures revealed a larger preference effect in the accurate condition than the random condition. Although the scores on the two measures were in the same direction on average for each condition, there was no significant correlation between the two scores after collapsing across the accuracy conditions [$r(45) = .26$, $p = .09$] or for

each accuracy condition considered separately [accurate: $r(19) = .04$ $p = .86$; random: $r(26) = -.03$, $p = .90$].

DISCUSSION

In experiment 1, the IAT reflected more than just exposure to associations; it was also sensitive to whether those associations were perceived as accurate. When participants knew in advance that the adjectives associated with the groups were arbitrary, as indicated by a coin flip, the relative preference for the group paired with positive adjectives was significantly reduced compared to when participants were instructed that the adjectives accurately described the groups. Furthermore, IAT scores were not affected by the number of association trials that the participants were exposed to, suggesting that association strength is not an important factor or at least a factor that diminishes in importance beyond 120 trials.

Even though both the implicit and explicit measures were affected in a similar way by the accuracy manipulation, there was no correlation between the implicit and explicit measures. However, this is not surprising given recent results demonstrating that individual IAT scores are contaminated by individual difference in executive functioning (Klauer et al., 2010). Perhaps because of this contamination, the magnitude of an individual's IAT effect was not predicted by the magnitude of an individual's responses on the explicit ratings. In other words, the covariance between the measures was weak even though perceived accuracy affected them similarly.

It is not surprising that the accuracy manipulation affected the explicit measures of attitude, but its effect on the IAT was unexpected. There was, however, one difference between the explicit and implicit measures: unlike the explicit ratings, there was a small preference effect for the IAT even in the random condition. Furthermore, there is a clear alternative interpretation of these results: participants in the random condition knew that the learned associations were arbitrary, and it is possible that they put less effort into the initial learning of the associations for this condition. If this occurred, the pattern of results merely reflects inattention in the random condition. To address this alternative interpretation, experiment 2 ensured that participants learned the associations even in the random condition.

EXPERIMENT 2

Experiment 2 was a replication of experiment 1 but with a memory test to ensure that participants paid attention and learned the associations during the learning task. The length of the learning task was not varied in this experiment because length had no effect for experiment 1. Similarly, IAT block order and the adjective/group associations were not included as nuisance variables because they produced no significant effects for experiment 1. This resulted in a 2 (accurate vs. random) $\times$ 2 (implicit measure vs. explicit measure) mixed design.

Participants

Forty-nine University of Maryland undergraduates, 35 women and 14 men, participated in the experiment. They received extra credit for their participation.

Procedure

The procedure of the second experiment was similar to that of the first experiment, with a few modifications. During the learning task, each Luupite and Niffite name was paired with a specific adjective from the appropriate list. For example, if the positive words were associated with the Luupites, the adjective "wonderful" might be specifically paired with the name "Neenolup." In the learning phase, each name was preceded only by the adjective with which it was paired.

Memory testing was embedded in the learning task by presenting a test question after every ninth trial of the learning task. Interleaved testing served two purposes. First, because testing was interleaved with the learning task, participants were made aware of which associations they had or had not learned, which motivated them to focus their attention on associations that required additional learning. Second, this interleaved method is commonly used in the memory experiments to study the "testing effect" (Karpicke & Roediger, 2008; Roediger & Karpicke, 2006b), which is the finding that retrieval practice promotes greater long-term retention as compared with additional study without testing. In testing effect experiments, repeated successful retrievals of the same information has been shown to promote accurate retention of the associations between novel word pairs with delays of days or weeks, particularly if the test trials include feedback as was the case in this experiment (see Roediger & Karpicke, 2006a for a review). Finally, we note that a post-IAT memory test would not necessarily demonstrate memory equivalent during the IAT. More specifically, the IAT may itself serve as an additional opportunity to rehearse the associations, and this would be particularly true for the accurate condition if there is more automatic evaluation of the groups in that condition.

The test questions were similar in format to the regular trials of the learning task. One of the adjectives from the sets was presented in the middle of the screen, and two names, both belonging to the appropriate group, appeared in the upper corners of the screen. One of the names was the name that was consistently paired with the adjective on the screen. Just as for a normal learning phase trial, the location of the names corresponded to the buttons assigned on the answer box. Participants attempted to identify which specific name the adjective was paired with by pushing the appropriate button. As with the regular learning trials, participants received accuracy feedback for these test trials. They were instructed that these trials were "test trials" to make sure they were learning the associations.

After completing the learning task, participants completed the explicit measure and then the Luupite/Niffite IAT. The adjective/group associations were fixed so that the Luupites were always associated with the positive words and the Niffites were always associated with the negative words for

all participants. The block order was fixed for all participants such that the compatible block was always the first of the mixed blocks. Other than these changes, the procedures and instructions were identical to those of the previous experiment.

RESULTS

Participants in the random condition learned the associations as effectively as the participants in the accurate condition, as revealed by an independent samples *t*-test comparing accuracy between these conditions [$t(49) = .08$, $p = .93$]. One-sample *t*-tests were conducted separately for each condition; learning was compared against the chance test value of 0.5 (50% accuracy). Participants in both the accurate and random condition performed above chance; accurate condition, precision: 71% [$t(21) = 27.32$, $p < .01$]; random condition, precision: 71% [$t(28) = 22.32$, $p < .01$]. Only four participants performed below 50% precision on the test questions. There was no qualitative difference in the results when these participants were excluded, and so all of the participants' data were used in the reported analyses. Demonstrating that participants progressively learned the associations, test accuracy was lower for the first two tests. This long-term learning trend rules out an explanation of memory performance based on short-term memory (Atkinson & Shiffrin, 1968; Peterson & Peterson, 1959). Average accuracy across both conditions was 80% when these first two trials were not included, demonstrating a high degree of learning.

Cronbach's alpha indicated that it was appropriate to combine the two questions of the explicit measure ($\alpha = .90$). The methods used to calculate the participants' explicit and IAT scores were identical to the methods used for experiment 1. Table 2a reports average IAT reaction times for each block, Table 2b reports the untransformed difference scores for both the IAT and explicit measures, and Table 2c reports the standardized versions of these difference scores.

As seen in Table 2b for the explicit measure, there was a greater preference for the positive group in the accurate condition than in the random condition [$t(47) = 7.56$ $p < .01$]. Similarly, for the IAT there was a greater preference for the positive group in the accurate condition than in the random condition [$t(47) = 2.50$, $p < .01$]. A one-sample *t*-test indicated that the average IAT effect in the random condition was not significantly different from zero [$t(27) = 1.65$, $p = .11$]. The average score on the explicit measure in the random condition was also not significantly different from zero [$t(27) = 1.01$, $p = .32$]. In the accurate condition, the average IAT effect was significantly different from zero [$t(20) = 3.82$, $p < .01$]. The average score on the explicit measure was also significantly different from zero [$t(20) = 10.98$, $p < .01$].

In summary, these results offer no evidence that the accuracy manipulation affected the two measures differently even though the degree of associative learning between adjectives and names was equal as demonstrated by the adjective–name memory test questions. Furthermore, unlike Experiment 1, there was no relative preference for the group paired with positive adjectives in the random condition according to both the explicit and implicit measures. Similar to experiment 1, there were no significant correlations between the two explicit and implicit measures after collapsing across accuracy [$r(49) = .19$, $p = .20$] or for each accuracy condition considered separately [accurate: $r(21) = -.19$, $p = .41$; random: $r(28) = -.02$, $p = .90$].

Table 2a. Experiment 2 IAT response times

	Block			
	Compatible		Incompatible	
Accurate condition	Mean (milliseconds)	Standard deviation (milliseconds)	Mean (milliseconds)	Standard deviation (milliseconds)
Accurate	818	203	986	282
Random	749	132	767	91

Table 2b. Experiment 2 untransformed differences

	Measure			
	IAT		Explicit	
Accurate condition	Mean (milliseconds)	Standard deviation (milliseconds)	Mean	Standard deviation
Accurate	167	209	3.76	1.57
Random	18	98	0.30	1.59

Table 2c. Experiment 2 standardized differences

	Measure			
	IAT		Explicit	
Accurate condition	Mean	Standard deviation	Mean	Standard deviation
Accurate	0.55	0.67	0.85	0.67
Random	0.15	0.47	−0.64	0.68

DISCUSSION

The results of this experiment were almost identical to those in the first experiment. Furthermore, the memory test demonstrated that the results cannot be explained by inattention to the associations during learning in the random condition. These findings provide additional evidence that IAT performance is sensitive to the perceived accuracy of newly learned associations.

Experiments 1 and 2 found evidence that although participants learned the associations in the random condition, they did not form an explicitly acknowledged attitude on the basis of these associations. This may have occurred if they encoded the adjective–name pairings as specific semantic associations without evaluating the groups on the

basis of these associations. To test this idea, experiment 3 manipulated accuracy in a different manner. The coin flip manipulation in experiments 1 and 2 made it clear that the adjective–name associations were arbitrary. However, an arbitrary association is not the same as an inaccurate association. To call something inaccurate implies that the opposite is true (i.e., negation). Indeed, in the accuracy manipulation of Gregg et al. (2006), after learning the adjective–name associations, participants were instructed that a computer error had reversed the adjective–name pairings (i.e., they were told that what they learned was inaccurate). In their study, this negation manipulation reversed the explicit measure but did not change the IAT effect. Thus, negated associations, rather than random associations, might be necessary to produce dissociations between the IAT and the explicit measures of attitude.

To test whether the IAT is sensitive to negated associations, experiment 3 used the modifier "not" before each adjective during the initial learning. On the basis of prior research, we had reason to expect that people would not automatically reverse the meanings of words when presented with the modifier "not." For instance, Gawronski and Bodenhausen (2006) proposed that implicit attitudes are insensitive to accuracy beliefs because the associations they are based on are immune to negation. Deutsch, Gawronski, and Strack (2006) found support for this claim, observing that simultaneous presentation of an affirming ("a") or negating ("no") word with a valenced word (e.g., "a party" or "no disease") produced a large effect for an explicit evaluation task of the valenced words but failed to affect the implicit evaluative priming produced by the valenced words for a subsequent target. A study by Mayo, Schul, and Burnstein (2004) provides another example in which people failed to incorporate negation into their implicit evaluation. They had people read character descriptions such as "Tom is a tidy person" (affirmation) or "Tom is not a tidy person" (semantic negation). Following these descriptions, they indicated whether statements such as "Tom's clothes are folded neatly in his closet" where congruent with the description. Mayo et al. found that people were faster to correctly respond to congruent than incongruent statements in the affirmation condition, but the opposite was true in the semantic negation condition. In other words, they were able to appreciate the explicit meaning of the negation manipulation and respond accurately, but their automatic evaluation of situation did not incorporate the negation, which produced response conflict that slowed down responses. This result is strikingly similar to the results of Gregg et al. (2006); even though participants in the Mayo et al. study explicitly knew the negated meaning (analogous to explicit attitudes in the Gregg et al. study), their reaction times indicated the opposite association (analogous to the IAT effect in the Gregg et al. study).

EXPERIMENT 3

In experiment 3, the accurate condition was identical to the accurate condition in the first two experiments. In the other condition, which we term the "negation" condition, the adjectives were presented with the negative modifier "not" during the learning task. We hypothesized that participants exposed to the negated adjectives might still form associations between the groups and those adjectives. If these associations were captured by the IAT, the IAT would show a preference for the group associated with positive adjectives, regardless of whether the adjectives were modified by "not." However, the participants' explicit attitudes should be reversed by the negations in the negation condition. Thus, the implicit and explicit measures should dissociate, with the negation manipulation only affecting performance on the explicit attitude measure. Alternatively, if the IAT is sensitive to the accuracy of associations at the time of learning as suggested by experiments 1 and 2, then both measures should reverse in the negation condition.

Participants

The participants were 29 University of Maryland undergraduates. Twenty-one women and eight men participated and received extra credit for their participation.

Procedure

The procedure of the experiment was identical to that of the first experiment except as noted.

The Learning Task

In the *negation* condition, all the adjectives in the learning task were preceded by the word "not." For example, if the Luupites were paired with the list of positive adjectives, a Luupite name might be preceded by "not wonderful" or "not fantastic," whereas a Niffite name might be preceded by "not horrible" or "not awful." In the *accurate* condition, the adjectives were presented unmodified. For example, if the Luupites were paired with the positive adjectives, a Luupite name might be preceded by the adjectives "wonderful" or "fantastic," whereas the Niffite names might be preceded by "horrible" or "awful." In both conditions, the participants were told that the descriptions of the groups were true.

Measures

The measures used in the third experiment were the same as in the previous experiments. It is important to note that the "nots" were omitted in the presentation of the IAT, and the data were scored according to adjectives' valences. Thus, if the "nots" reversed the valence of the adjectives, this would result in a negative IAT score. For example, when Luupite names were paired with "not fantastic" in the learning phase, the block of the IAT in which Luupite names and positive words (e.g., fantastic) were assigned to the same response key was designated the compatible block. The explicit measure was scored the same way: in the negation condition, the group paired with the negative adjectives was considered the negative group. Conversely, even if those adjectives were preceded by the word "not,"

the group paired with the positive adjectives was considered the positive group.

RESULTS

A Cronbach's alpha indicated that it was acceptable to combine the two explicit questions ($\alpha = .89$). Table 3a reports average IAT reaction times for each block, Table 3b reports the untransformed difference scores for both the IAT and explicit measures, and Table 3c reports the standardized versions of these difference scores.

In the accurate condition, participants had a greater preference for the group associated with the positive adjectives, for the explicit measure, $t(12) = 3.10$, $p < .01$, but in the negation condition, when the adjectives were preceded by the negative modifiers, they showed a preference for the group associated with the (negated) negative adjectives. The difference in preference between these two conditions was significant [$t(27) = 7.24$, $p < .01$].

For the IAT, there was a greater preference for the group associated with the positive adjectives in the accurate condition. [$t(12) = 3.75$, $p < .01$]. This preference reversed in the negation condition such that there was a stronger preference for the group associated with the negative adjectives [$t(15) = 4.32$, $p < .01$].

In summary, both the implicit and explicit measures were sensitive to the reversed meaning of the adjectives (when they were negated), and the pattern of responses was similar on the two measures. Unlike experiments 1 and 2, there was a significant correlation between the explicit and implicit measures after collapsing across accuracy [$r(29) = .55$, $p < .01$]. However, this correlation was likely due to the accuracy manipulation (i.e., a two-point correlation depending on accuracy condition), and, similar to experiments 1 and 2, there was no correlation when each accuracy condition was considered separately [accurate condition: $r(13) = -.18$, $p = .57$; negation condition: $r(16) = -.04$, $p = .87$].

Table 3a. Experiment 3 IAT response times

	Block			
	Compatible		Incompatible	
Accurate condition	Mean (milliseconds)	Standard deviation (milliseconds)	Mean (milliseconds)	Standard deviation (milliseconds)
Accurate	782	163	940	220
Negation	771	129	690	88

Table 3b. Experiment 3 untransformed differences

	Measure			
	IAT		Explicit	
Accurate condition	Mean (milliseconds)	Standard deviation (milliseconds)	Mean	Standard deviation
Accurate	159	167	3.04	3.54
Negation	−80	87	−4.56	2.05

Table 3c. Experiment 3 standardized differences

	Measure			
	IAT		Explicit	
Accurate condition	Mean	Standard deviation	Mean	Standard deviation
Accurate	0.36	0.35	0.89	0.75
Negation	−0.24	0.23	−0.72	0.43

DISCUSSION

The explicit measures of attitude demonstrated that participants reversed their attitudes with negated associations. However, this was also true for the IAT. When the Luupites were paired with the negated positive adjectives in the learning phase, the IAT revealed a preference for the Niffites. These results were surprising in light of past research demonstrating that people have a difficult time negating associations on the basis of the modifier "not." This experiment expanded on the results of experiments 1 and 2 by demonstrating that, whereas newly learned arbitrary (random) associations produced small or absent IAT effects, newly learned inaccurate (negated) associations produced a reversed IAT effect.

GENERAL DISCUSSION

In a series of three experiments, we tested whether the IAT is sensitive to the perceived accuracy of newly learned associations. Prior research with newly learned associations found that accuracy manipulations can affect explicit attitude measures while leaving IAT effects unchanged (Gregg et al., 2006). However, unlike that study, our study presented the accuracy manipulation in advance of learning to rule out the possibility that this difference is due to a primacy effect. Furthermore, unlike that study, our study only used a single IAT test to eliminate possible contamination from learning during prior IAT performance. Other research with newly learned associations found that the mere supposition of an arbitrary association can create an IAT effect (De Houwer, 2006). However, unlike that study, our study included two conditions of perceived accuracy to assess whether IAT effects are greater when associations are supposedly accurate. Across all three experiments, we found clear evidence that the IAT is sensitive to the perceived accuracy of newly learned associations.

In experiment 1, participants learned to associate novel groups with sets of positive or negative adjectives. Some

participants were told that the adjectives accurately described the groups, whereas for other participants the adjectives were randomly assigned as indicated by a coin flip. After participants learned the associations, they completed an IAT and gave explicit ratings of their attitudes towards the groups. When they were instructed that the associations were accurate, both the IAT and their explicit ratings indicated a preference for the group that was paired with positive adjectives. However, when these pairings were determined by a coin flip, this preference was reduced for both measures. One possible explanation of smaller preferences for the coin flip condition was that participants did not attend to the associations. To address this possibility, experiment 2 replicated the results of experiment 1 while including a memory test of the newly learned associations. This memory test was interleaved with learning to promote greater attention during learning and strong long-lasting memories based on retrieval practice (i.e., the "testing effect"). Performance on this memory test was the same for both conditions, demonstrating equivalent learning, but nevertheless the IAT and explicit measures were nearly identical to the results of experiment 1. However, a randomly determined association is not necessarily the same as learning that an association is inaccurate (i.e., reversed). Therefore, in experiment 3, the adjectives were presented with the negative modifier "not" for one of the conditions. As with experiments 1 and 2, both the IAT and the explicit attitude ratings were similarly affected by the accuracy manipulation, although in this case, the negation condition fully reversed preferences. Thus, across all three experiments, the IAT reflected the perceived accuracy of the newly learned associations.

Implications for Attitude Models

These findings have important theoretical implications depending on whether one subscribes to the theory that implicit attitudes are separate from explicit attitudes (i.e., a dual attitude model) and whether the IAT is considered to be a valid measure of attitude.

First, we consider implications of our results under a classic single attitude model. From this perspective, the observed consistency between IAT results and explicit ratings is sensible, particularly if the IAT is deemed a valid attitude measure. Fazio and Olson (2003) make a distinction between implicit and explicit *measures* of attitude, which can be distinguished even if one subscribes to a single attitude model; the IAT is an implicit measure of attitude, which means that the method by which it assesses attitude is not immediately obvious. Implicit measures of attitude are relatively free from presentation issues (i.e., the desire to be socially acceptable, such as not appearing racist in modern society). More generally, if respondents withhold their true attitude when giving an explicit rating because of self-presentation concerns, this can explain why the IAT and explicit attitude measures produce different results in certain circumstances. However, in the current situation, there was no reason for participants to withhold their true attitude during the explicit rating task because the IAT was testing attitudes for novel groups. Therefore, if the IAT is a valid measure of attitude and if there is just one kind of attitude, our accuracy manipulations should have affected both measures in a similar manner, as was observed.

Next, we consider dual attitude models (e.g., Wilson et al., 2000). According to this perspective, explicit attitudes are separate from implicit attitudes, and measurement of implicit attitudes requires an implicit measure such as the IAT. Gawronski and Bodenhausen (2006) argued that explicit attitudes are "evaluative judgments that are based on syllogistic inferences derived from any kind of propositional information that is considered relevant for a given judgment," (Gawronski & Bodenhausen, 2006, p. 694) whereas implicit attitudes are associative and consist of automatic affective reactions tied to the relevant groups. This implies that the associative processes underlying implicit attitudes cannot be altered by propositional processes. Gawronski and Bodenhausen proposed that implicit attitudes are immune to "truth values," which means that they cannot be reversed or neutralized by the respondents' explicit beliefs. From this perspective, the sensitivity of the IAT to the perceived accuracy of the newly learned associations is unexpected, particularly so for experiment 3, which was designed to test the role of propositional processes. Alternatively, it could be that the current findings highlight an exception to implicit attitudes' immunity to propositional information; it may be that the initial formation of associations can be influenced by propositional processes, but these associations become immune to propositional information thereafter.

Conclusions

Putting aside the single versus dual attitude debate, these experiments demonstrate that the IAT does not merely reflect learned associations. If learned associations are known to be random or inaccurate at the time of learning, then the associations will not produce an IAT effect and might even produce a reversed IAT effect. Thus, in contrast to the claims of Karpinski and Hilton (2001), the IAT does not just reflect environmental associations; it is also necessary that the environmental associations are believed to be valid at the time that they are learned. By analogy, if an individual is only exposed to a racial stereotype in situations where the stereotype is presented as inaccurate, then this environmental association should not produce an IAT effect. However, considering the results of Gregg et al. (2006), the same cannot be said if the accuracy of a stereotype is questioned after it has been learned. When considering whether to use the IAT, our results indicate the importance of initial exposure to associations and the manner in which those associations are learned.

ACKNOWLEDGEMENT

This research was supported by National Science Foundation Grant BCS-0843773 to David E. Huber.

REFERENCES

Anderson, N. H., & Barrios, A. A. (1961). Primacy effects in personality impression-formation. *Journal of Abnormal and Social Psychology*, *63*(2), 346–350.

Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. In K. W. Spence, & J. T. Spence (Eds.), *The psychology of learning and motivation* (Vol. 2). London: Academic Press.

Banse, R., Seise, J., & Zerbes, N. (2001). Implicit attitudes towards homosexuality: Reliability, validity, and controllability of the IAT. *Zeitschrift Fur Experimentelle Psychologie*, *48*(2), 145–160.

Blair, I. V., Ma, J. E., & Lenton, A. P. (2001). Imagining stereotypes away: The moderation of implicit stereotypes through mental imagery. *Journal of Personality and Social Psychology*, *81*(5), 828–841.

Blanton, H., & Jaccard, J. (2006). Arbitrary metrics in psychology. *American Psychologist*, *61*(1), 27–41.

Brendl, C. M., Markman, A. B., & Messner, C. (2001). How do indirect measures of evaluation work? Evaluating the inference of prejudice in the Implicit Association Test. *Journal of Personality and Social Psychology*, *81*(5), 760–773.

Dasgupta, N., & Greenwald, A. G. (2001). On the malleability of automatic attitudes: Combating automatic prejudice with images of admired and disliked individuals. *Journal of Personality and Social Psychology*, *81*(5), 800–814.

De Houwer, J. (2006). Using the Implicit Association Test does not rule out an impact of conscious propositional knowledge on evaluative conditioning. *Learning and Motivation*, *37*(2), 176–187.

Deutsch, R., Gawronski, B., & Strack, F. (2006). At the boundaries of automaticity: Negation as reflective operation. *Journal of Personality and Social Psychology*, *91*(3), 385–405.

Eagly, A. H., & Chaiken, S. (1998). Attitude structure and function. In D. T. Gilbert, & D. T. Fiske (Eds.), *The handbook of social psychology* (4th edn, Vol. *1*, pp. 269–322). New York: Oxford University Press.

Fazio, R. H., Han, H. A., & Olson, M. A. (2006). The influence of experimentally created extrapersonal associations on the Implicit Association Test. *Journal of Experimental Social Psychology*, *42*(3), 259–272.

Fazio, R. H., & Olson, M. A. (2003). Implicit measures in social cognition research: Their meaning and use. *Annual Review of Psychology*, *54*, 297–327.

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

Gawronski, B., & Strack, F. (2004). On the propositional nature of cognitive consistency: Dissonance changes explicit, but not implicit attitudes. *Journal of Experimental Social Psychology*, *40*(4), 535–542.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The Implicit Association Test. *Journal of Personality and Social Psychology*, *74*(6), 1464–1480.

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

Greenwald, A. G., Nosek, B. A., & Sriram, N. (2006). Consequential validity of the Implicit Association Test: Comment on Blanton and Jaccard (2006). *American Psychologist*, *61*(1), 56–61.

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

Gregg, A. P., Seibt, B., & Banaji, M. R. (2006). Easier done than undone: Asymmetry in the malleability of implicit preferences. *Journal of Personality and Social Psychology*, *90*(1), 1–20.

Han, H. A., Czellar, S., Olson, M. A., & Fazio, R. H. (2010). Malleability of attitudes or malleability of the IAT? *Journal of Experimental Social Psychology*, *46*(2), 286–298.

Karpicke, J. D., & Roediger, H. L. (2008). The critical importance of retrieval for learning. *Science*, *319*(5865), 966–968.

Karpinski, A., & Hilton, J. L. (2001). Attitudes and the Implicit Association Test. *Journal of Personality and Social Psychology*, *81*(5), 774–788.

Klauer, K. C., Schmitz, F., Teige-Mocigemba, S., & Voss, A. (2010). Understanding the role of executive control in the Implicit Association Test: Why flexible people have small IAT effects. *Quarterly Journal of Experimental Psychology*, *63*(3), 595–619.

Mayo, R., Schul, Y., & Burnstein, E. (2004). "I am not guilty" vs "I am innocent": Successful negation may depend on the schema used for its encoding. *Journal of Experimental Social Psychology*, *40*(4), 433–449.

Olson, M. A., & Fazio, R. H. (2004). Reducing the influence of extrapersonal associations on the Implicit Association Test: Personalizing the IAT. *Journal of Personality and Social Psychology*, *86*(5), 653–667.

Peterson, L. R., & Peterson, M. J. (1959). Short-term retention of individual verbal items. *Journal of Experimental Psychology*, *58*(3), 193–198.

Roediger, H. L., & Karpicke, J. D. (2006a). The power of testing memory basic research and implications for educational practice. *Perspectives on Psychological Science*, *1*(3), 181–210.

Roediger, H. L., & Karpicke, J. D. (2006b). Test-enhanced learning—taking memory tests improves long-term retention. *Psychological Science*, *17*(3), 249–255.

Rothermund, K., & Wentura, D. (2004). Underlying processes in the Implicit Association Test: Dissociating salience from associations. *Journal of Experimental Psychology. General*, *133*(2), 139–165.

Wilson, T. D., Lindsey, S., & Schooler, T. Y. (2000). A model of dual attitudes. *Psychological Review*, *107*(1), 101–126.