Journal of Personality and Social Psychology
2001, Vol. 81, No. 5, 760–773

0022-3514/01/$5.00 DOI: 10.1037//0022-3514.81.5.760

How Do Indirect Measures of Evaluation Work? Evaluating the Inference of Prejudice in the Implicit Association Test

C. Miguel Brendl
INSEAD, Fontainebleau

Arthur B. Markman
University of Texas at Austin

Claude Messner
University of Heidelberg

There has been significant interest in indirect measures of attitudes like the Implicit Association Test (IAT), presumably because of the possibility of uncovering implicit prejudices. The authors derived a set of qualitative predictions for people's performance in the IAT on the basis of random walk models. These were supported in 3 experiments comparing clearly positive or negative categories to nonwords. They also provided evidence that participants shift their response criterion when doing the IAT. Because of these criterion shifts, a response pattern in the IAT can have multiple causes. Thus, it is not possible to infer a single cause (such as prejudice) from IAT results. A surprising additional result was that nonwords were treated as though they were evaluated more negatively than obviously negative items like insects, suggesting that low familiarity items may generate the pattern of data previously interpreted as evidence for implicit prejudice.

What do you think of flowers? Would you evaluate them positively? If so, what do you think of Larnists? Do you think a Larnist is more negative or more positive than a flower? As you may have realized, a *Larnist* is not an English word; we made it up. Presumably, you do not have a prestored opinion of Larnists, and so you would have to compute it ad hoc.

Pretend for a moment that Larnists are people who live in houses with odd-numbered addresses. If we provided you with evidence that you are less likely to associate Larnists than flowers with pleasant words, would you conclude that you are prejudiced against people who live in odd-numbered houses? We suggest not, because you probably have no prior evaluation of Larnists (i.e., people who live in odd-numbered houses) in general. It would seem safer to conclude that something other than a prior evaluation of Larnists must have produced your behavior.

In this article, we explore recent research on the indirect measurement of individual differences in attitudes. This work claims to uncover implicit prejudices that may not be consciously accessible to the people who hold them. We focus our discussion on measures that are based on response competition, specifically on the Implicit Association Test (IAT) presented by Greenwald, McGhee, and Schwartz (1998), which is the best developed measure of implicit evaluations. We begin by discussing definitions of prejudice to know what these tests aim to measure. Then, we describe the IAT in detail. Next, we present qualitative predictions for people's performance on this task derived from random walk models. Finally, we test these predictions, and discuss the implications of this work for indirect measures of attitudes.

Prejudice and Discrimination

Part of the excitement surrounding indirect measurements of attitudes is the possibility that researchers may discover hidden prejudices. These prejudices may be unavailable to more explicit measures either because people do not want to reveal their attitudes or because they do not even recognize that they have them. To place our discussion in this article in context, it is important to distinguish between prejudice and discrimination. There is a long history of research on these topics, and a complete review of the use of these terms is beyond the scope of the present article. Prejudice is defined as involving an evaluative prejudgment (attitude) toward some group (typically an out-group). That is, prejudice necessarily involves an affective reaction and a prestored judgment. This affective reaction is generally thought of as being negative (e.g., Allport, 1954; Brown, 1995; Dovidio & Gaertner,

C. Miguel Brendl, INSEAD, Fontainebleau, France; Arthur B. Markman, Department of Psychology, University of Texas at Austin; Claude Messner, Department of Psychology, University of Heidelberg, Heidelberg, Germany.

We thank Tony Greenwald, Jamie Pennebaker, Bill Swann, Joachim Vosgerau, and four anonymous reviewers for helpful comments on drafts of this article. We also thank Irena Ebert, Nils Kaltenbach, Elfrun Sophr, and Almut Stromberger for serving as experimenters. This work was supported by German Science Foundation Grant DFG BR1722/1–2 and an INSEAD Research and Development award, both to C. Miguel Brendl, by National Science Foundation Grant SBR-9905013 to Arthur B. Markman, and by a Transcoop award from the German American Academic Council to C. Miguel Brendl and Arthur B. Markman.

Correspondence concerning this article should be addressed to C. Miguel Brendl, INSEAD, Boulevard de Constance, 77305 Fontainebleau Cedex, France, or to Arthur B. Markman, Department of Psychology, University of Texas, Mezes Hall 330, Austin, Texas 78712. Electronic mail may be sent to miguel.brendl@insead.edu or to markman@psy.utexas.edu.

1986; Ehrlich, 1973; Fiske, 1998; Gardner, 1994; Newcomb, Turner, & Converse, 1965; Sherif & Sherif, 1956), although a small minority of authors have suggested extending the definition to cover positive feelings as well (e.g., "all Italians cook well"; Klineberg, 1954; Secord & Backman, 1964; Simpson & Yinger, 1985). On either definition, if someone has positive attitudes toward two groups, but likes one group more than another, they are not negatively prejudiced against the lesser liked group. Neither in everyday language nor in the prejudice literature is less "love" equated with "hate." When prejudice is viewed as racism, it is clear that the standard definition of prejudice is meant, that is, negative, derogatory feelings.

Discussions of prejudice lead naturally to discussions of discrimination. Discrimination necessarily involves an *overt behavior* that "unjustifiably" favors one group over another, but there need not be any affect involved (Brewer, 1994; Fiske, 1998; Jones, 1972; Reid & Holland, 1997).[1] Often, when one is (negatively) prejudiced against a group, it is assumed that this prejudice will manifest itself in discriminatory behavior against that group. However, as Dovidio and Gaertner (1986) explained, "prejudice does not always lead to discrimination and . . . discrimination may have causes other than prejudice" (p. 3). Two important points must be made. First, indirect measures of attitudes cannot assess discrimination, because discrimination is overt behavior, not an attitude. Thus, even if a reliable and valid measure of implicit prejudice were developed, additional research would have to establish the link between these measured attitudes and behavior. Second, even if such a link were established, not all attitudes that lead to discrimination should necessarily be called prejudice. Prejudice is typically defined in terms of a negative attitude. It would be possible, however, for a person to intentionally treat one group better than another even if that person has strong positive attitudes toward those groups.

The IAT

The IAT is an ingenious use of response competition designed to measure attitudes indirectly (Greenwald et al., 1998). In this test, people respond to two classes of stimulus words. The first class of words is one that can be easily evaluated along an *attitudinal dimension*. In the IAT, this class consists of a set of obviously pleasant words and another set of obviously unpleasant words (in which pleasant–unpleasant is the attitudinal dimension). People respond by pressing one key when they see an obviously pleasant word (e.g., peace), and a second key when they see an obviously unpleasant word (e.g., ugly).

In addition to the attitudinal sets, there are two *target sets*, which are the words of interest in the IAT. For example, an IAT examining racial attitudes might contain stereotypically White names (e.g., Betsy, John), and stereotypically Black names (e.g., Temeka, Rasaan) as its two target sets. In some blocks of the IAT, people respond to the White names with the identical key as the pleasant words and the Black names with the identical key as the unpleasant words. On other blocks, the response mappings are switched, so the Black names and pleasant words get one key and the White names and unpleasant words get the other (see Figure 1).

What would happen if research participants had an implicit evaluation of White people as good and Black people as bad? Greenwald et al. (1998) suggested that participants should be faster in responding when the White names get the same key as pleasant words than when the White names get the same key as unpleasant words. In this case, responses to the Black names should be faster when they are paired with unpleasant words than when they are paired with pleasant words. This pattern of data is indeed obtained with White participants.

Greenwald et al. (1998) interpreted this pattern of data in two ways. First, they suggested that the IAT measures relative preference, that is, the degree to which White names are more associated with pleasant terms (or perhaps pleasant valence) than Black names. This interpretation alone leaves open the possibilities that Black names are associated with positive affect, just to a lesser degree than White names, or alternatively that they are associated with negative affect. Second, these investigators interpret their IAT effects as "indicating the pervasiveness of unconscious forms of prejudice" (p. 1475) and as implicit racism (p. 1476). The terms *racism* and *prejudice* imply an association of Black names with negative affect, particularly because people are responding faster to the pairing of Black names with unpleasant items than with pleasant attitudinal items.

To facilitate the rest of the discussion, we introduce some terminology. As mentioned above, the key objects being evaluated by the IAT (e.g., the sets of White and Black names in the example above) are the target sets. The pleasant and unpleasant words that are not part of the target sets are called *attitudinal sets*. The assignment of the attitudinal sets to keys results in one "pleasant" and one "unpleasant" key. Blocks of trials in which the target sets are mapped to keys that are consistent with their expected (implicit) valence are called *compatible blocks* (see Figure 1a). These blocks are expected to lead to relatively fast responses. Blocks in which the target sets are mapped to response keys that are inconsistent with their expected valence are called *incompatible blocks* (see Figure 1b). These blocks are expected to lead to relatively slow responses. The difference in response time between the incompatible and the compatible block is called the *IAT effect*. When the IAT effect is significantly greater than 0, it is taken as evidence for an implicit prejudice or implicit racism against the group (e.g., Black names) paired with pleasant words in the incompatible block. In other words, an association of Black names with negative valence is assumed to contribute to the IAT effect.

Interpreting the IAT Results

So far, we have given just an intuitive description of why the IAT effect would be obtained when people evaluate one target set more positively than the other. It is worth exploring this prediction in more detail. To this end, we discuss the IAT in terms of random walk models of response times, which have emerged as important models of performance in response-time tasks (e.g., Nosofsky & Palmeri, 1997; Ratcliff, 1988; Townsend & Ashby, 1983). Random walk models (and their close cousins, diffusion models) have been used successfully to explore speed of responding in a number of tasks like recognition memory, classification, and automaticity.

Figure 2 depicts a basic random walk for a task in which people must respond if an item is pleasant or unpleasant. The x axis on

[1] It is beyond the scope of this article to discuss when differential treatments between groups are unjustified.

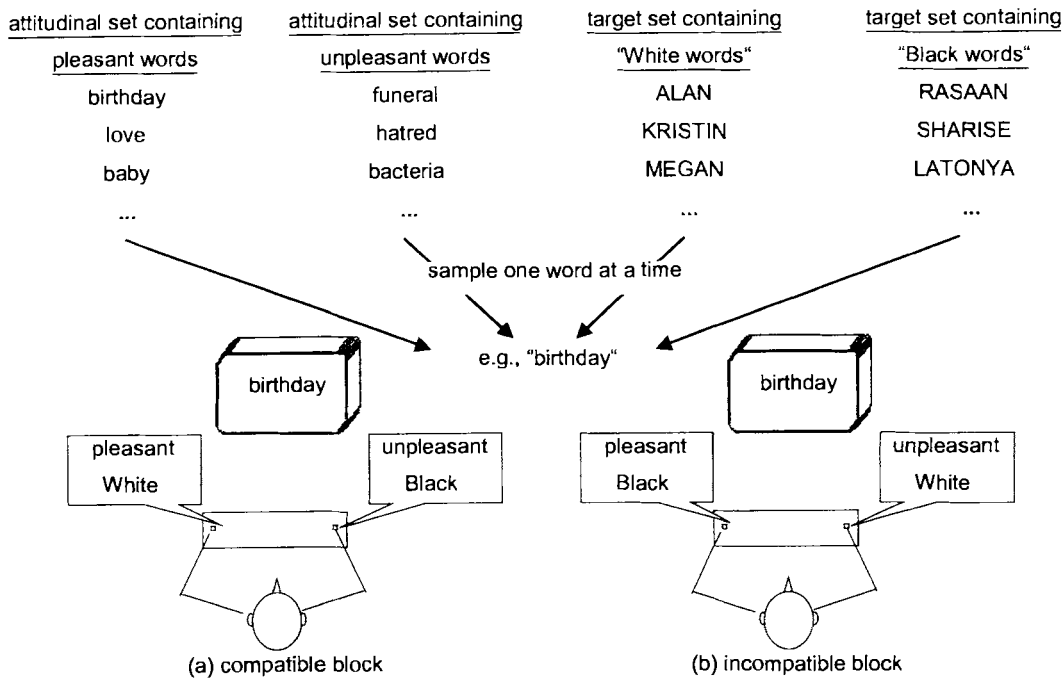

Figure 1. Illustration of the IAT's main blocks (i.e., Blocks d and g). One word at a time is sampled randomly from the four word sets and presented on the computer screen (e.g., *birthday*). The participant has to categorize the word as belonging to one out of four categories using two response keys. The meaning of the response keys differs in the (a) compatible block and (b) incompatible block.

this graph represents time, and the y axis represents the relative amount of pleasant–unpleasant information accumulated on a trial. The model starts at Time 0 at a neutral point. At each time step, the stimulus is sampled for properties pertaining to the response, and a step is taken in the direction of the response consistent with the information obtained. For example, if the word *peace* (a pleasant word) were presented, then the information sampled about this word would generally be pleasant information. On each time step a new piece of evidence that *peace* is a pleasant word is obtained, reflected in the upward movement of the line in the graph. When one of the response thresholds is reached (the two response thresholds are shown as bold lines in Figure 2), the response associated with that threshold is executed. In this example, pleasant information about the stimulus is extracted until the pleasant key threshold is reached (the bold line above the x axis) and then the response "pleasant" is executed. The amount of time required to make a response is proportional to the number of time steps required to reach a response threshold.

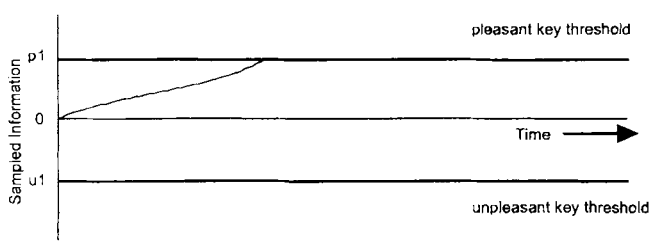

Figure 2. Simple illustration of a random walk. Time is shown along the *x*-axis, and information accumulation along the *y*-axis. The response thresholds are shown with bold lines above and below the *x*-axis. A response is executed when information accumulates to either the pleasant key threshold (p1) or the unpleasant key threshold (u1).

To apply a random walk model to the IAT, we assume that when people sample the stimulus word, they obtain two kinds of information: valence and identity. The valence information (pleasant–unpleasant) is required to determine whether the pleasant or unpleasant key should be pressed. The identity information is used to determine whether a word is a member of the target sets. This assumption leads to one important difference between the target set words and the attitudinal words. For the target set words (e.g., the White name *Mary*), both identity and valence information contribute to a response, but for the attitudinal words (e.g., *peace*) only valence information moves the current position toward one of the thresholds (see Figure 3). We examine the assumption that target words and attitudinal words are treated asymmetrically in more detail in the General Discussion.

Identity information does not affect responses to words in the attitudinal sets, because these words are selected such that their superordinate categories are nondiagnostic of the target sets. For example, the target word has both valence information (pleasant) and identity information (White name) that are related to the response keys in this task. In contrast, for the attitudinal word *peace* it is only its valence information (pleasant) that is related to the responses. The identity information of the superordinate categories of *peace* (e.g., *emotional state*) is not useful for making a response (though the valence of the superordinate category may be assessed).

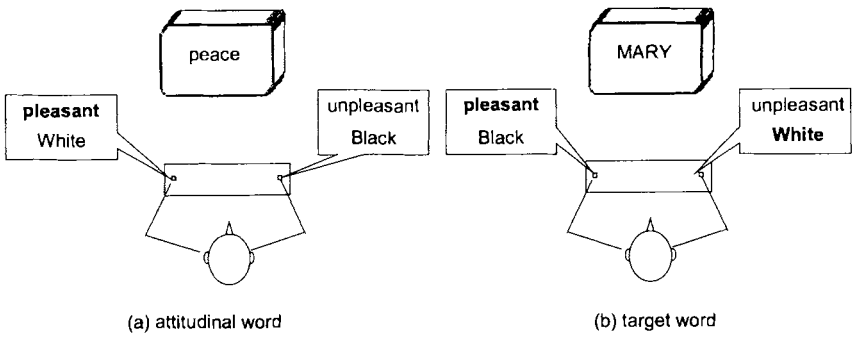

Figure 3. Schematic depiction of the asymmetry in response conflict between responding to attitudinal words and target words in an incompatible block. Bold entries symbolize responses that seem tempting from the participant's perspective. (a) In response to the word *peace* only the *pleasant* label seems tempting. (b) In response to the White name *Mary* the labels *White* and *pleasant* seem tempting. There is only response competition in b.

We discuss two versions of the random walk model. In the first model, we assume that the pleasant and unpleasant response thresholds remain fixed across blocks. Later we explore what happens when different thresholds are set for different blocks. For expository purposes, we talk about an example in which the target sets are flowers (a clearly pleasant set) and insects (a clearly unpleasant set; cf. Greenwald et al., 1998, Experiment 1). To conserve space, we frame the predictions of this model for compatible and incompatible blocks in terms of responding to the pleasant attitudinal set and the target sets.

Random Walk Model With Fixed Thresholds

There are four central cases for this model. First, in the compatible block when a pleasant attitudinal word is presented (e.g., *peace*), the model should behave as in Figure 2. At each time step, the valence information obtained pushes the model toward the pleasant–flower response threshold. Any identity information obtained will not affect the response (except insofar as that identity information itself is pleasant). When the threshold is reached the pleasant response is executed. This response occurs relatively quickly.

Second, in the compatible block of trials when a target word (e.g., *rose*) is presented, the sampled information will consist of some positive valence information and some identity information. All of this information will push the model toward the pleasant–flower threshold. Thus, a fairly rapid response should be made in this case as well. This case is illustrated in the top panel of Figure 4.

The next two cases involve the incompatible block in which one key is used to respond to pleasant attitudinal words (e.g., *peace*) and insects while the other key is used to respond to the unpleasant attitudinal words (e.g., *war*) and the flowers. When a pleasant attitudinal word is presented, the sampled information moves the model toward the pleasant–insect threshold as before. Because there is no information about pleasant attitudinal words that pushes the model toward the unpleasant response threshold, a fairly rapid response should be made in this condition as well. This case is illustrated in the middle panel of Figure 4.

The final case is the presentation of a target word (e.g., *wasp*) in the incompatible block. This case involves competition among the responses. The valence information sampled pushes the model toward the unpleasant–flower response threshold, but the identity

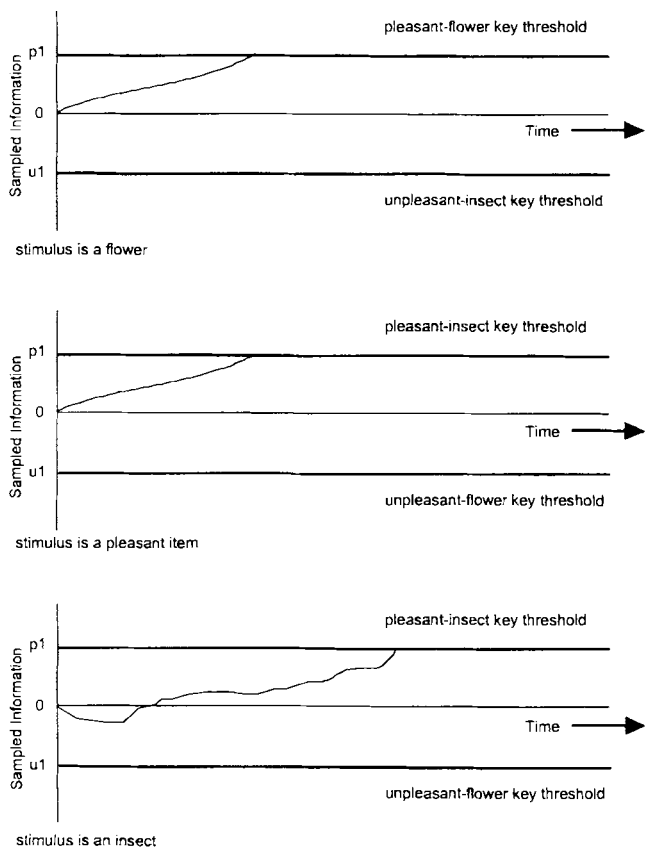

Figure 4. Three additional cases of a random walk model with a fixed threshold. The top panel shows responses to the target set in the compatible block. The bottom two panels show the responses to the pleasant and target set items in the incompatible block.

information pushes the model toward the pleasant–insect threshold. Eventually, the model should reach the proper threshold and respond with the pleasant–insect key. These responses will take longer to make than those in the compatible block, because more information must be sampled before a response threshold is reached than was required in the compatible block. In addition, errors are more likely in this block than in the other block. This case is illustrated in the bottom panel of Figure 4.

To summarize, this random walk model predicts that the response latency to the pleasant and unpleasant attitudinal items should be the same in the compatible and incompatible blocks. In contrast, responses to the target set items should be faster in the compatible block than in the incompatible block (see left half of Table 1). A pattern of data like this would be strong evidence for an implicit positive attitude toward one of the target-set groups and an implicit negative attitude toward the other target-set group.

Modeling a Criterion Shift: Random Walk Model With Variable Thresholds

Using the random walk model, it is also possible to show that some patterns of data in the IAT, which are quite similar to the one described in the previous section, would not provide strong evidence for implicit prejudice. In particular, we can explore patterns of data that would occur if people changed their response criterion as a function of the perceived difficulty of the block of trials (e.g., Ratcliff, 1988).[2] In the General Discussion we report evidence showing that participants in the IAT perceive the incompatible block to be more difficult than the compatible block. A response criterion shift is modeled as a movement of the response threshold.[3] Making the criterion more conservative involves shifting the threshold further from the starting point so that more information must be sampled before a response is executed.

Figure 5 shows a simple case: In Figures 2 and 4, the pleasant key thresholds p1 (i.e., pleasant 1) and the unpleasant key threshold u1 (i.e., unpleasant 1) were used. Note that participants are highly motivated to avoid error because they receive instant feedback in the case of an error and were instructed to minimize errors. If participants perceive that the task is more difficult in one block than it had been in a previous block, they may choose to move their thresholds further from the neutral point (say to pleasant 2 and u2) to decrease the likelihood of making an error in that block. In this case, the response times to pleasant items in the incompatible block would be longer than the response times to the pleasant items in the compatible block. Thus, finding longer response times for the pleasant and unpleasant items in the incompatible block than in the compatible block would suggest that people are setting a more conservative criterion for responding to all words in the incompatible block than in the compatible block (see Table 1).

If people use a different threshold to respond in the compatible block than they do in the incompatible block, it becomes difficult to interpret the IAT as a measure of implicit prejudice as discussed below. To make the reason for this assertion clear, we return to the example like the one we described at the beginning of the article in which one target set is pleasant (e.g., flowers) and the other is novel (e.g., nonwords such as *gize*). Like the fictitious Larnists described above, nonwords are items for which people should have no precomputed valence (although it is possible that these items have valence that is computed "on the fly").

We assume that if a new block is more difficult than one performed previously, then participants set a more conservative criterion for responding in that block (i.e., they move their threshold further from the starting point). Similarly, if they recognize that the task is easier, they move their threshold closer to the starting point. Consider the compatible block first. Responses to pleasant and unpleasant attitudinal words should be fairly fast, because the less conservative threshold is used. Responses to pleasant target words (e.g., flowers) are fast, because both the valence information and the identity information push responses toward the pleasant–flower threshold. Finally, responses to the nonwords should be relatively fast, because identity information moves the model toward the unpleasant–nonword threshold, and there is no valence information associated with these neutral stimuli.

Now we turn to the incompatible block. Here the response competition to the valenced target set makes some trials in this block more difficult. Thus, we expected participants to use a more conservative criterion in the incompatible block than in the compatible block. This analysis predicts that responses to all of the items in the incompatible block will be slower than they were in the compatible block. Of course, slower responses to *both* target sets in the incompatible block than in the compatible block is the definition of an IAT effect.[4] According to Greenwald et al. (1998) this pattern of data indicates implicit prejudice. Note, however, that we began the example by assuming that the nonword items were neutral. Thus, the interpretation of the IAT would not reflect the actual valence of the items.

Table 1
Response Latencies Predicted From Random Walk Models With Fixed Versus Variable Thresholds

	Fixed threshold		Variable threshold	
Block	Attitudinal word[a]	Target word	Attitudinal word[a]	Target word
Compatible	fast	fast	fast	fast
Incompatible	*fast*	slow	*slow*	slow

Note. Italic entries highlight predictions that are different for a fixed versus variable threshold model. In the incompatible block without a threshold shift only reactions to target words are slow, but with a threshold shift reactions to all words (target words and attitudinal words) are slow.
[a] Pleasant or unpleasant.

[2] The prevalence of criterion shifts in different blocks of response time tasks has led to the development of more complex response-time techniques such as Ratcliff and McKoon's (1989) variable response-deadline procedure.

[3] We use the term *criterion* to denote the psychological-response criterion and the term *threshold* to denote the amount of information required in the random walk model to make a response.

[4] Greenwald et al. (1998) defined the IAT effect as the slowing of responses across all trials in the incompatible compared with the compatible block.

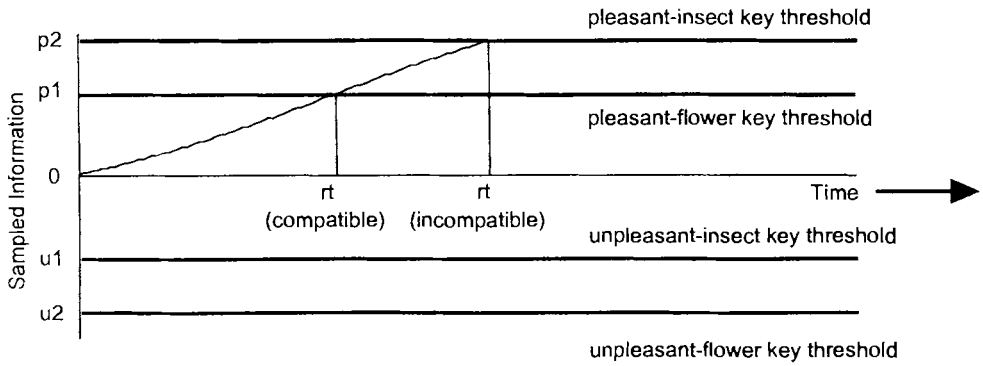

Figure 5. An example of changes in response time (rt) caused by making the response threshold more conservative by moving it from p1 to p2. The abbreviation p1 is the pleasant–flower key threshold in the compatible block and p2 is the pleasant–insect key threshold in the incompatible block; u1 is the unpleasant–insect key threshold in the compatible block and u2 is the unpleasant–flower key threshold in the incompatible block.

Two Conditional Probabilities

The previous example reveals a potential problem with using the IAT as a test for implicit prejudice. There are two conditional probabilities that must be high for any test to be a valid indicator of the construct being assessed. First, it must be shown that given certain situations in which the construct holds, there is a high probability of obtaining particular patterns of data. In the case of the IAT diagnosing prejudice, this means demonstrating that there is a response-time difference between the compatible and incompatible blocks when one target set is clearly negative (e.g., names of a prejudiced group) and the other target set is less negative or even positive. Greenwald et al. (1998, Experiment 1) provided a strong demonstration of this proposition in a study with insects and flowers as target sets.

Second, the reverse conditional probability must also be high. That is, the presence of an IAT effect (a difference in response time between the compatible and incompatible block) must be a reliable marker of prejudice. As we demonstrate, the threshold shift described in the previous section calls this aspect of the IAT into question. For example, it is possible to observe a difference in response time between the compatible and incompatible blocks even when the target set does not have a prestored valence. In sum, whereas the probability of obtaining an IAT effect given prejudice is high, it is not clear that there is a high probability of there being prejudice given an IAT effect.

As a demonstration of this point, imagine a person who has a strong positive feeling toward Caucasians, but has no attitude toward African Americans, is run in the Black-name–White-name IAT. In the incompatible block, this person will recognize that something about responding to the White names is difficult, and will thus set a more conservative criterion for responding than in the compatible block. This more conservative criterion will lead to slower responses to all items in the incompatible block than in the compatible block. Thus, we might be tempted to infer that this participant is prejudiced against African Americans despite the absence of any attitude at all.

It is, of course, difficult to imagine someone who has no attitude at all toward other people. In the following section, we discuss three experiments that explore this issue. Each of these studies uses the IAT method and compares a target set of known valence with a set of nonwords (which presumably have no prestored valence). The first two studies compare the nonwords against a target set known to be positive (White names) and against one known to be negative (insects). In the third study, another version of the test of nonwords against insects is discussed briefly as it rules out a possible alternative explanation for the results of the studies.

Experiments: The IAT and Nonwords

Method

Overview and Design

Each participant performed an IAT designed to assess attitudes to the two target categories. One study used the target sets White names and nonwords (Experiment 1) and the other two used the target sets insects and nonwords (Experiments 2 and 3). For the associated attitudinal attribute dimension, participants had to discriminate pleasant from unpleasant attitudinal words. The experiment consisted of seven blocks: (a) the target category discrimination (insect names vs. nonwords), (b) the evaluative (attitudinal) attribute discrimination (pleasant vs. unpleasant attitudinal words), (c) a practice block with the combination of Tasks a and b, (d) a data block with the combination of Tasks a and b in which the data were counted, (e) the target category discrimination with reversed mappings of words to keys, (f) a practice block with the combination of Tasks b and e, and (g) a data block with the combination of the Tasks b and e. The main dependent measure is the reaction time to the target items in Blocks d and g.[5] In addition, we examined the response times to the pleasant and unpleasant attitudinal words in Blocks d and g. The order in which the keys were associated with words was counterbalanced between participants.

Materials and Apparatus

We used the Farnham Implicit Association Test for Windows (1998, Version 2.2) software made available through the IAT World Wide Web

[5] The results are substantially the same if the results from the practice blocks are also examined. However, a plot of the response times reveals large nonlinear practice effects in the first block with each response pairing. Thus, we report only analyses from the second block of each type.

site at the University of Washington.[6] The experiment was run under Windows 95 on two Pentium computers with 14-in. (approximately 36 cm) CRT monitors set up at a viewing distance of about 60 cm.

In the White name versus nonword version of the studies (Experiment 1), the participants were native speakers of English. The White names were 18 names drawn from the set used by Greenwald et al. (1998). The nonwords were created by changing one or two letters of the names. Americans who were native speakers of English were then shown these items to determine whether these words reminded them of any English words. Only items that did not remind them of any English words were used in this study. The items used in this study are shown in the Appendix.

The two insect versus nonword studies (Experiments 2 and 3) were run with native speakers of German. German stimulus words from four categories (pleasant, unpleasant, insects, nonwords) were used (n = 16 per category). Pleasant and unpleasant words were taken from Hager and Hasselhorn (1994) and from Fazio, Sanbonmatsu, Powell, and Kardes's (1986) word list provided in Bargh, Chaiken, Govender, and Pratto (1992). Insect names were taken from Greenwald et al. (1998). On the basis of pilot tests, we tried to eliminate names of insects uncommon in German. Further, to create the most negative set of insect names, the least negative insects (e.g., bee) were deleted. German sounding nonwords were created from the German insect names by changing two or three letters in each word (see the Appendix). In some cases a single letter was added to or deleted from the nonword. Pilot tests showed that it was nearly impossible to recognize the original insects and that the nonwords were not associated with other German words.

Participants

Participants in the White-name–nonword study (Experiment 1) were 20 native speakers of English recruited from tourist areas in Heidelberg in return for a bottle of sparkling wine or a 5 DEM coupon for ice cream. Participants in the first insect–nonword study (Experiment 2) were 32 native speakers of German, recruited from a pedestrian zone in a busy shopping district of Heidelberg in return for a bottle of sparkling wine. One person was excluded from this study after the experiment because for one word category in one block of trials all of his responses were wrong. Finally, participants in the second insects–nonword study (Experiment 3) were 52 native speakers of German from the same population. Two participants were excluded because they made an extraordinary amount of errors (36% and 83%).

Procedure

Structure of blocks of trials. Trials in all experiments were presented using the seven blocks described above. Participants were instructed to respond as quickly and accurately as possible by pressing one of two keys marked with green stickers. One key was located on the left side of the keyboard, the other one on the right side. Each block started with instructions that assigned either 1 or 2 category labels to each response key depending on the type of block. Category labels were *pleasant*, *unpleasant*, and *insects* in all experiments and the addition of *White* in Experiment 1, *artificial words* in Experiment 2, and a somewhat different way of introducing nonwords in Experiment 3, to be discussed in more detail below.

Each trial began by displaying a category word (either target category or pleasant–unpleasant category) centered horizontally and vertically on the screen until the correct response key was pressed. During the whole block the category labels were placed to the left or right of the category word corresponding to their assignment to the left or right response key. When two category labels were assigned to each response key, one label was printed above the other one on each side of the category word. During the whole block a green circle was visible underneath the category word indicating correct responding. However, when the wrong response key was pressed, the green circle was replaced with a red x. Together with the category word this x remained on the screen until the correct response key was pressed at which point the category word was erased and the green circle replaced the red x. After a delay of 250 ms the next trial started. Each block ended with a summary feedback on reaction time and percent correct responses. Category words were drawn randomly without replacement from the respective word lists.

Types of blocks. Each practice block (i.e., Blocks a, b, c, e, f) in the White-name–nonword study (*Experiment 1*) consisted of 36 trials. In Practice Blocks c and f, of the 18 stimuli in each category, 9 were drawn randomly. Each main block (i.e., Blocks d and g) consisted of 72 trials; thus, all of the 18 stimuli in each category were presented once. In Experiments 2 and 3 we reduced the 18 stimuli per category to 16. Therefore each practice block consisted 32 trials and each main block consisted of 64 trials.

Results and Discussion

Data Reduction

We analyzed the data from the two blocks that combined all word types (Blocks d and g from the description above). To be consistent with Greenwald et al. (1998) we excluded the first two trials of each of these blocks because of their high variability. Because of the usual deviations of reaction time data from normality, we log-transformed the reaction times. However, for clarity the means are reported untransformed.

In all three studies, incorrect responses were excluded, as were trials with response times below 300 ms and above 3,000 ms. This method differs slightly from the one used by Greenwald et al. (1998). They set short response times to 300 ms and very long response times to 3,000 ms. Because very short or very long reaction times probably reflect responses involving different processes than the ones we are interested in, we chose to exclude them. We analyzed our studies using Greenwald et al.'s method as well, and the results were substantially the same.

The error rates in all of the studies were quite low. On average, fewer than 5% of all trials were errors. Conditions with longer response times tended to have slightly higher error rates. We do not discuss the error data further in this article.

Experiment 1: White Names and Nonwords Study

The mean response times for each item type and block are shown in Table 2. Two aspects of these data are striking. First, response times in the compatible block (in which the White names were paired with the pleasant key) were significantly shorter overall (M = 783) than were response times in the incompatible block (in which the White names were paired with the unpleasant key; M = 1,112), $t(19) = 11.18$, $p < .05$. Thus, there was a significant IAT effect. The standard interpretation of this effect is that White names are relatively preferred to nonwords.

This effect can also be seen by looking only at the target sets. Participants categorized White names faster in the compatible block (M = 761 ms) than in the incompatible block (M = 1,096 ms), $t(19) = 9.27$, $p < .05$. Likewise, they categorized nonwords faster in the compatible block (M = 829 ms) than in the incompatible block (M = 1,222 ms), $t(19) = 9.25$, $p < .05$.

[6] This site has since moved to http://buster.cs.yale.edu/implicit/index.html.

Table 2
Mean Response Times (in Milliseconds) for Each Item Type and Condition of Experiment 1

Block	Item type			
	White names	Nonwords	Pleasant	Unpleasant
Compatible	761	829	759	794
Incompatible	1,096	1,222	1,062	1,108

Note. For all item types, the difference between the mean response times in the compatible and incompatible blocks was statistically significant at $p < .05$ by t test.

These data also provide evidence for a threshold shift across blocks. An examination of the pleasant attitudinal words reveals faster response times to these items in the compatible block (M = 759 ms) than in the incompatible block (M = 1,062 ms), $t(19) = 9.36$, $p < .05$. Similarly, responses to the unpleasant attitudinal words were also faster in the compatible block (M = 794 ms) than in the incompatible block (M = 1,108), $t(19) = 6.99$, $p < .05$.

IAT effects like the one obtained here have previously been interpreted as evidence for individual differences in implicit prejudice (in the present study, against nonwords). There are two reasons to dispute this interpretation. First, the nonwords were designed to be items for which people did not have any prior associations, yet prejudice implies a preexisting evaluation. Second, the longer response times to the pleasant and unpleasant items in the incompatible block than to those in the compatible block provides evidence of a change in response threshold across blocks. These data suggest that people simply recognized that something about the incompatible block was more difficult than the compatible block. Given evidence for a criterion shift, the response times to the nonwords could well be a reflection of this criterion shift rather than of them having negative valence. We now turn to a study using insects (an obviously negative category) and nonwords as target sets. Given the logic described above, we predicted faster response times when the insects are paired with unpleasant and the nonwords with pleasant than when the insects are paired with pleasant and the nonwords with unpleasant; that is, nonwords would look as if they were positive.

Experiment 2: Insects and Nonwords Experiment

We examined the response times and error rates from an IAT performed with insects and nonwords as target sets. We began with an overall test of the IAT effect (defined as the overall response time to all items in the insects–pleasant block and in the insects–unpleasant block). Surprisingly, there was a significant IAT effect in the opposite direction than was expected. People were significantly slower in the insects–unpleasant block (M = 955 ms) than in the insects–pleasant block (M = 840 ms), $t(30) = 3.99$, $p < .05$.

Next, we conducted t tests of response times in each block type separately for trials presenting insect names and nonwords. Participants categorized nonwords faster when combined with unpleasant words (M = 888 ms) than with pleasant words (M = 1,026 ms), $t(30) = 3.93$, $p < .05$. Similarly they categorized insects faster when combined with pleasant words (M = 879 ms) than with unpleasant words (M = 986 ms), $t(30) = 2.61$, $p < .05$ (see Table 3). These data are surprising in light of our expectation that people would be faster when insects were paired with unpleasant words than when they were paired with pleasant words. We examine this issue in more detail in the General Discussion.[7]

As in Study 1, we analyzed the response times to the pleasant and unpleasant words. The data from this study show evidence consistent with a threshold shift across blocks. Responses to the pleasant items were faster in the insect–pleasant block (M = 778 ms) than in the insect–unpleasant block (M = 867 ms), $t(30) = 3.27$, $p < .05$. Similarly, responses to the unpleasant items were faster in the insect–pleasant block (M = 812 ms) than in the insect–unpleasant block (M = 939 ms), $t(30) = 4.37$, $p < .05$. Thus, once again there is evidence that people recognized that one block was more difficult than the other and slowed down all of their responses accordingly.

So far, we have assumed that people had no prior evaluation of the nonwords, and thus that the presence of an IAT effect cannot be interpreted as a reflection of an implicit prejudice. One possible explanation for the effects in these studies is that the nonwords reminded people of other words that had negative valence. We tried to construct the materials so that the nonwords would have no associations, but we might have failed.

To examine this possibility, we showed each participant the nonwords after they completed the IAT and asked them to generate a real word associated with it. Then, they rated the valence of the nonwords, the associates to the nonwords, and the insects on a Likert scale ranging from 1 (*Negative*) to 6 (*Positive*).[8] Both the nonwords (M = 3.11) and the associations generated for these words (M = 3.33) were given significantly more positive ratings than were the insects (M = 2.48), $t(31) = 4.78$, and $t(29) = 5.41$, both $p < .05$.[9] This study suggests that the results of the present study did not arise because of the valence of the nonwords. We discuss below the possibility that the implicit valence of positively rated nonwords could be negative. However, the present results are difficult to reconcile with a position that the nonwords reminded participants of other negative words.

Independent of our conclusions about nonwords, we have to explain why the clearly negative category of insects is paired more quickly with pleasant than unpleasant attitudinal words in this IAT. The random walk model with fixed thresholds predicts that when two negative target categories are combined in one IAT, each one of them is categorized faster when paired with unpleasant than with pleasant words. This model is inconsistent with the data.

The model assuming variable thresholds provides a better account of the data. It follows from this model that when a pairing of a target category (i.e., nonwords) with an attitudinal category (i.e.,

[7] We also ran a version of the insect name versus nonword version of the IAT with native speakers of English and observed the same pattern of data.

[8] Participants were not able to come up with associates for all of the nonwords. Obviously, participants could not rate the valence of the associate in this case, but they still rated the valence of the nonword and the insect.

[9] The degrees of freedom for these t tests differ because there were two participants who were unable to come up with associates for any of the nonwords.

Table 3
Mean Response Times (in Milliseconds) for the Insects/Nonwords Study (Experiment 2)

Block	Item type			
	Insects	Nonwords	Pleasant	Unpleasant
Insects with pleasant words	879	888	778	813
Insects with unpleasant words	986	1,026	867	939

Note. For all item types, the difference between the mean response times in the compatible and incompatible blocks was statistically significant at $p < .05$ by t test.

unpleasant) is perceived as difficult, the responses to all items in that block are slowed, including those to the other target set (i.e., insects), which happens to be paired with the pleasant attitudinal category. This slowing of all trials in the incompatible block explains why the clearly negative category of insects is paired more quickly with pleasant than with unpleasant attitudinal words. We must point out that although we can account for the way nonwords and insects can appear to have a negative and positive valence in the IAT, respectively, the difficulty of pairing nonwords with pleasant words is not predicted by the random walk model. We return to this issue in the General Discussion.

Experiment 3: Target Set Items and Labels

Another potential alternative explanation for the observed behavior with the nonwords is that people have a negative reaction to the category label *artificial words*. On this view, they responded to nonwords as they would to negative items because they did not like the category label.[10] To explore this possibility, we ran another version of the insect–nonword study in which we created a cover story designed to get people to treat the nonwords and the label for the nonwords positively. If the participants in this study still showed an IAT effect in which the nonwords are treated more negatively than the insects, then we could conclude that something other than the prestored valence of the items is contributing to the IAT effect.

Like the previous insects–nonwords study, this experiment was run in German. A translation of the cover story is as follows:

> In international kindergartens it is often observed that the children understand each other even if they do not speak the same language. This phenomenon is typically explained as the result of nonverbal behavior. We are interested in studying the influence of the words themselves. In particular, there is some evidence that you will automatically think positive thoughts when you hear positive words of a language you do not know.
>
> To explore this idea, we are going to show you some positive words in the language Lositio. Lositio was the language of a highly developed European culture that was located in a small geographic area. The most important thing for this study is that you be unfamiliar with Lositio. If you have already heard about this language or are familiar with some of the words of the language, please contact the experimenter. Otherwise, we will begin the experiment. All of the words that you do not know are positive words in Lositio.
>
> By the way, you might be interested to know that the word Lositio also had a meaning in the language: It means "joy."

The cover story was effective in giving people a positive impression of the category label *Lositio*. Participants were asked to rate the goodness of all of the words in this study (including the label *Lositio*) after completing the IAT on a scale ranging from 1 (*Negative*) to 6 (*Positive*). The mean valence rating for this category label was 4.94, which was significantly above the midpoint of the scale, $t(49) = 7.83$, $p < .05$.

The response time data from this study are shown in Table 4. As can be seen from this table, the pattern of data is substantially the same as that obtained in the previous studies. The t tests between the compatible and incompatible blocks reveal significant differences for all four types of stimuli.[11] Once again, the responses to the pleasant and unpleasant items are slower in the incompatible block than in the compatible block, which is indicative of a criterion shift. Also, as before, people were slower to respond when the same key was used to respond to insects and unpleasant words (and nonwords and pleasant words) than when they had to respond with the same key to insects and pleasant words (and nonwords and unpleasant words).

General Discussion

The present studies show that each one of the following people would be diagnosed as negatively prejudiced against African Americans in the Black-name–White-name IAT: A person with (a) prestored negative evaluations of Black names, (b) prestored positive evaluations of White names without evaluative associations to Black names, (c) stronger prestored positive evaluations of White than Black names without negative evaluations of either, and (d) low familiarity of Black names in the absence of any prestored evaluation of Black names. In our view there is little doubt that according to commonly agreed on definitions of prej-

[10] We thank one anonymous reviewer for bringing up this possibility to us and suggesting Study 3.

[11] An ANOVA on these data revealed an unexpected Block (compatible vs. incompatible) × Category (insect vs. nonword vs. pleasant vs. unpleasant) × Order (compatible first vs. incompatible first) interaction, $F(1, 48) = 31.66$, $p < .05$, and a Block × Category interaction, $F(1, 48) = 5.58$, $p < .05$. Of the eight mean comparisons between compatible and incompatible blocks (two orders for four word categories) in this design, two could be regarded as problematic for our analysis. First, in the order condition in which insects were paired with unpleasant words in the first block, there was no reliable difference between the response time to the unpleasant words in the incompatible block ($M = 932$ ms) and in the compatible block ($M = 954$ ms), $t < 1$. In contrast, when insects were paired with pleasant words in the first block, responses to the unpleasant items were faster in the compatible block ($M = 848$ ms) than in the incompatible block ($M = 973$ ms), $t(23) = 4.1$, $p < .05$. A second mean comparison did not quite reach significance; when insects were paired with pleasant words in the first block, responses to the nonword items were nonsignificantly faster in the compatible block ($M = 953$ ms) than in the incompatible block ($M = 1{,}006$ ms), $t(23) = 1.8$, $p = .08$. In contrast, when insects were paired with unpleasant words in the first block, responses to the nonword items were significantly faster in the compatible block ($M = 971$ ms) than in the incompatible block ($M = 1{,}140$ ms), $t(25) = 4.3$, $p < .05$. All other items (i.e., six of the eight mean comparisons) showed faster response times in the compatible block than in the incompatible block for both orders. Because order did not interact with IAT effects in any of the other four experiments we have done, we do not discuss this discrepant finding further.

Table 4
Mean Response Times (in Milliseconds) for the Insects/Artificial Language Study (Experiment 3)

Block	Item type			
	Insects	Nonwords	Pleasant	Unpleasant
Insects with pleasant words	918	962	803	901
Insects with unpleasant words	1,022	1,073	902	952

Note. For all item types, the difference between the mean response times in the compatible and incompatible blocks was statistically significant at $p < .05$ by t test.

udice only the first person would be called prejudiced. Whereas prejudice leads to an IAT effect, an IAT effect does not unambiguously indicate prejudice, because it can have multiple causes.

Our main hypothesis is that people use a different response criterion in different blocks of the IAT, which compromises the interpretation of this test as a measure of individual differences in implicit prejudice. The data from the three studies are consistent with a random walk model with variable response thresholds. In particular, the IAT effect reflects slower responses to all items, even to the pleasant and unpleasant attitudinal words. As discussed above, when there is a threshold shift between blocks a central potential problem for interpreting the IAT is that a target set that has no valence (or even one that has positive valence) can exhibit a pattern of response times that has often been interpreted as reflecting implicit (negative) prejudice.

These studies also revealed an unexpected pattern of data with nonwords. Response times were slower in the insects–pleasant/nonwords–unpleasant condition than in the nonwords–pleasant/insects–unpleasant condition (Experiments 2 and 3). In the random walk model, we assumed that people only use valence and identity information when determining their response. This model predicted that the insects–pleasant condition should have been slower than the insects–unpleasant condition because of response competition between pressing the unpleasant key and the insect key.

Target Set and Attitudinal Set Asymmetry

A key assumption of the random walk model is that the responses to the attitudinal set items are influenced only by their valence, but the responses to the target set items are influenced by both their valence and their category membership. This asymmetry contrasts with a prominent intuition that there is a symmetric association between the target and attitudinal sets that have the same valence. One might think that a symmetric connection would predict the observed data that responses to the attitudinal items in the inconsistent condition would be slower than the responses to those items in the consistent condition.[12]

If one adopts an associative approach (as opposed to the random walk model that we presented in the introduction), this model would not predict the observed pattern of data because of the *fan effect* (Anderson, 1983). The fan effect has been incorporated into spreading activation models to account for the observation that concepts with fewer associations activate the concepts to which they are connected more strongly than do concepts with many connections. The fan effect is typically implemented by dividing the amount of activation sent from a node in the network by the number of connections leaving that node. In the context of the IAT, an associative model would assume that the concept *White name* would have a connection to the concept *pleasant*. The target word *Mary* (a White name) would activate *pleasant* more strongly than the attitudinal word *peace* (a pleasant item) would activate *White name*, because there are more concepts connected to *pleasant* in memory than are connected to *White name*.

In addition, in discussions of the IAT, the target words are assumed to be connected directly both to the superordinate target category and to the attitudinal dimension. In contrast, the attitudinal words only have direct connections to the attitudinal dimension. They have at most an indirect connection to the target category (through the attitudinal dimension). Thus, for example, *Mary* is connected both to *White name* and to *pleasant*, but *peace* is connected only to *pleasant*. To the extent that *peace* activates the target set *White names* it can only happen indirectly by first activating *pleasant*. Thus, the assumption that there is a symmetric association between the target set and the attitudinal set does not appear tenable.[13]

The Role of Familiarity of an Item

In this section, we consider three explanations for why responses to nonwords were slower when paired with pleasant words than with unpleasant words: (a) Familiarity of an item may affect performance on the IAT without mediation by valence, (b) unfamiliar items may have a prestored negative valence, and (c) the familiarity of an item may factor into an on-line computation of valence. It is difficult to rule out the first explanation, but because it would compromise the IAT as a measure of implicit valence, we focus on Possibilities 2 and 3 in this discussion.

Could participants actually have had a prestored implicit negative valence for unfamiliar items? It is hard to rule out this explanation conclusively, but there is some reason to doubt this possibility. First, ratings of goodness of insects and nonwords revealed that people's explicit ratings of the valence of insects was significantly lower than their ratings of the valence of nonwords. Second, in the last study we presented, we told people that the novel words were positive, gave them a reason to expect that they would react as if the words were positive, and also told them that the category label (the language *Lositio*) was a positive word in that language. Despite these efforts, people were still faster to respond in the insects–pleasant condition than in the insects–unpleasant condition.

The final possibility is that familiarity information influences computations of valence that take place when the item is presented. On this view, unfamiliar items have negative implicit valence. As indirect support for this possibility, there are some findings demonstrating that familiarity can lead to positive explicit valence. For example, in the mere exposure effect, prior

[12] We thank two reviewers for pointing out this possibility.

[13] If a spreading activation model that includes the assumptions in this section is generated, it actually predicts a pattern of data that differs substantially from what was observed. In particular, this model predicts a strong interaction between response compatibility and set type. Further details about this model are available from the authors.

experience with neutral or positive stimuli increases liking of those items (Bornstein & D'Agostino, 1992; Zajonc, 1968). Similarly, in the endowment effect, people who are given possession of an item value it more highly simply because it has been given to them (Kahneman, Knetsch, & Thaler, 1991). Yet, prior experience with negative stimuli appears to decrease liking of those items (Klinger & Greenwald, 1994). Klinger and Greenwald explain these mere-exposure experiments in terms of a misattribution of perceived familiarity of an item to its valence. It is conceivable that unfamiliarity is attributed to dislike in our case. Indeed, Mandler (1982) suggests that perceptions that are incongruent with existing schemata cause negative affect, and unfamiliar stimuli are likely to be schema inconsistent (see also Meyers-Levy & Tybout, 1989). Note, however, that all of these results have been obtained with explicit ratings of valence, and our explicit ratings resulted in a positive valence for nonwords. What exactly familiarity does to the IAT effect appears quite unclear currently.

The idea that low familiarity produces an on-line computation of negative valence would muddy the interpretation of the IAT as a measure of individual differences in attitudes, such as prejudices. As discussed above, prejudice assumes that people have a prestored negative attitude about some group. To the extent that negative attitudes are generated on the fly on the basis of familiarity, it is hard to argue that these attitudes reflect prejudice, because they are not prestored. Greenwald et al. (1998) clearly recognized these interpretational implications when asking, "Does the IAT measure implicit attitude, or is it an artifact of amount of exposure to the stimuli used to represent target concepts?" (p. 1477) and when identifying familiarity as a "possible alternative to the implicit racism interpretation" (p. 1476).

Consistent with the importance of familiarity as a potential alternative explanation of IAT effects, Greenwald and his colleagues (Dasgupta, McGhee, Greenwald, & Banaji, 2000; Greenwald et al., 1998) and others who have used the IAT (e.g., Ottaway, Hayden, & Oakes, 2001) have explored this issue. Tests have been performed in which items (e.g., names or faces) are equated for their familiarity. Under these circumstances, it has been demonstrated that there is a reliable IAT effect.

As discussed above, if the IAT is to be used as a test of implicit prejudice, then two aspects of its validity must be assessed. First, it must be demonstrated that when there is a difference in implicit valence there is a high probability of obtaining an IAT effect. Studies like the ones described in the previous paragraph have addressed this issue. Second, it must be demonstrated that when an IAT effect is obtained, it is highly likely that there is a difference in implicit prestored valence between the target sets and not "just" in familiarity or valence based on familiarity. This latter conditional probability has not been explored sufficiently. Indeed, the results obtained in our studies suggest that differences in familiarity may also cause a reliable IAT effect. This interpretation of our results suggests that the presence of an IAT effect does not necessarily reflect a difference in implicit prestored valence. Thus, no matter whether one believes that low familiarity affects the IAT directly or through on-the-fly valence, the influence of familiarity compromises the IAT as an unambiguous marker of prestored valences such as implicit prejudice.

What Causes Criterion Shifts?

In our presentation of the random walk model we examined predictions both with and without a threshold shift. These two versions of the model give qualitatively different predictions, and the data suggest that participants shifted their response criterion between the compatible and the incompatible block. Independent of the familiarity issue, the presence of a criterion shift by itself compromises the interpretation of the IAT as a test of implicit prejudice.

We have not firmly established the mechanism that leads to this criterion shift. If people's attitudes toward the target sets are not consciously available, then valence information cannot be used to set a criterion. However, there are other aspects of the task that people could use to determine the difficulty of a block. One possible mechanism is that they may recognize a propensity to make errors in the incompatible block, and thus shift their criterion to roughly equate the error rates across blocks. It is hard to provide evidence for this possibility, because if people successfully shift their criterion to equate error rates, then there will be no reliable difference in the error rates between the compatible and the incompatible blocks. It is difficult to use a null result as support for a theoretical claim.

A second possibility is that people have a subjective experience of difficulty caused by the response competition in the incompatible block (cf. Figure 3). On this view, the criterion shift occurs when people sense that a response is more difficult to make. To assess this possibility, we replicated Greenwald et al.'s (1998) Experiment 1, an IAT with insects and flowers as the target set, and asked people to rate the difficulty of each block immediately after completing that block on a scale from 1 (*not at all difficult*) to 6 (*very difficult*). Participants did not receive any performance feedback after the blocks. They rated the incompatible block as considerably more difficult ($M = 3.0$) than the compatible block ($M = 1.5$), $t(19) = 5.45$, $p < .05$. Finally, there is a strong positive correlation between size of the difference in the difficulty ratings and the size of the response time difference between the compatible and incompatible blocks, $r(19) = .61$, $p < .01$. These data support the hypothesis that participants shifted their response criterion as a result of perceiving the incompatible block as a more difficult task than the compatible block.

The Importance of Computational Models

One issue that emerges from this discussion is the importance of applying some kind of information-processing model to tasks designed to measure implicit associations. The predictions for people's performance in the IAT have previously been made on the basis of intuitively plausible assumptions. However, when a more specific information-processing model is applied to people's performance, some unforeseen conclusions can be drawn.

When we applied a random walk model to the IAT, it became clear that some patterns of performance would be consistent with a global shift in people's criterion for responding, which is modeled as a movement of the response threshold. Because of the criterion shift, it is possible to obtain a pattern of data indicative of implicit prejudice against a target set in the absence of a prestored negative valence for the items in that set.

Recently, there has been an increasing reliance on quantitative and computational models in social cognition (e.g., Read & Miller,

1998; Smith, 1996, 1998). This development may lead to a revival of theories that were abandoned because of a lack of specificity. For example, Read, Vanman, and Miller (1997) discussed how Gestalt theories in social psychology waned when it became difficult to generate predictions for them. They suggest that using parallel constraint satisfaction models (a variety of connectionist models) may permit a revival of these theories, because predictions can now be generated on the basis of computational models.

Similarly, social psychologists have long been interested in indirect measures. Projective techniques such as the Thematic Apperception Test (Murray, 1943) or related tests (e.g., Haire, 1950) were popular for a while but interest in them waned, presumably because of the difficulty of interpreting these tests. We hope that by examining measures like the IAT using information-processing models, we may avoid a similar fate for the tests now being developed.

Methodological Issues

The popularity of the IAT attests to the importance of developing unobtrusive measures of attitudes, particularly those attitudes that people may not be willing or able to report. The present studies suggest that, whereas the IAT is an important advance in our ability to measure implicit attitudes, it needs to be modified. There are two central problems with the IAT, and both of them arise because this test involves a pair of target sets. The first problem is that responses to a target set may be slowed as a result of a criterion shift made in response to difficulties processing the other target set. Because we have discussed this issue at length already, we do not say more about it here.

The second problem with the IAT is that it is difficult to interpret the results. The test itself gives at best a relative measure of one target set against another. However, in contradiction to this constraint of relativity, the results of the IAT are often interpreted as reflecting an implicit prejudice for one group over another. The problem with this interpretation is that (as discussed in the introduction) prejudice connotes a negative attitude toward a group. When two groups are measured relative to one another, it is possible for one group to be preferred to another without the second group being evaluated negatively. Adhering to an interpretation of relative evaluation requires refraining from assigning absolute valences (i.e., neither positive nor negative) as in the case of "prejudice" and "racism." Of course one could redefine prejudice and racism as reflecting purely relative preferences. If any perception of a difference in valence between the two groups were treated as a prejudice, then we gloss over those effects that arise as a result of negative valence toward a target group, and very likely the term *prejudice* would apply to our perception of almost any group.

Although the previous analysis assumes the IAT measures relative preference of two categories, it is not self-evident that merely putting two target categories into one IAT achieves this goal. As demonstrated by our analysis using the random walk model, the IAT does not lead to a comparison of the two target categories, but rather reflects the ease of retrieving evaluative information, identity information, and perhaps other information such as familiarity. For this test to be a marker of relative preference, there would have to be a single preference scale that is used to evaluate all items. Given the problems with common currency (e.g., utility) models of preference uncovered in the decision making literature, it is unlikely that there is a single preference scale against which all objects are measured (see Medin, Goldstone, & Markman, 1995, for a discussion of this issue).

These problems with relative measures of valence suggest that the development of new techniques for indirect measurement of attitudes should focus on ways of evaluating one target set at a time. In principle, a test could be constructed by having people evaluate only a single target set relative to some evaluative dimension. This type of test would eliminate problems in interpretation that arise because one target set is evaluated relative to another, although it would not solve the problem that there may be many factors that influence response time (such as familiarity). As we have discussed, further research must focus on the probability that an IAT effect signals a difference in implicit evaluation rather than some other factor.

Finally, the research on indirect measurements of attitudes has focused primarily on the development of techniques. The extensive amount of work required to develop the IAT reflects the subtlety necessary to develop an effective test. Nonetheless, there is a potentially dangerous circularity in these techniques. Once a test has been established, it is difficult to know whether a new effect in that test is a reflection of a difference in implicit valence or a difference in some other factor, because implicit valences by definition cannot be measured explicitly. Thus, it is difficult to get an independent assessment of the valence of stimuli used in an experiment. To break out of this circularity, it is necessary to explore the relationship between these measures of attitude and other behaviors. As discussed in the introduction, a distinction is often made between prejudice (an attitude) and discrimination (a behavior). Finding relationships between attitudes and behaviors is the next crucial step in the development of measures like the IAT.

References

Allport, G. (1954). *The nature of prejudice.* Boston: Beacon Press.

Anderson, J. R. (1983). A spreading activation theory of memory. *Journal of Verbal Learning and Verbal Behavior, 22,* 261–295.

Bargh, J. A., Chaiken, S., Govender, R., & Pratto, F. (1992). The generality of the automatic activation effect. *Journal of Personality and Social Psychology, 62,* 893–912.

Bornstein, R. F., & D'Agostino, P. R. (1992). Stimulus recognition and the mere exposure effect. *Journal of Personality and Social Psychology, 63,* 545–552.

Brewer, M. B. (1994). The social psychology of prejudice: Getting it all together. In M. P. Zanna & J. M. Olsen (Eds.), *The Psychology of Prejudice: The Ontario Symposium* (Vol. 7, pp. 315–329). Hillsdale, NJ: Erlbaum.

Brown, R. (1995). *Prejudice: Its social psychology.* Oxford, England: Blackwell.

Dasgupta, N., McGhee, D. E., Greenwald, A. G., & Banaji, M. R. (2000). Automatic preference for White Americans: Eliminating the familiarity explanation. *Journal of Experimental Social Psychology, 36,* 316–328.

Dovidio, J. F., & Gaertner, S. L. (1986). Prejudice, discrimination, and racism: Historical trends and contemporary approaches. In J. F. Dovidio & S. L. Gaertner (Eds.), *Prejudice, discrimination, and racism* (pp. 1–34). Orlando, FL: Academic Press.

Ehrlich, H. J., (1973). *The social psychology of prejudice: A systematic theoretical review and propositional inventory of the American social psychological study of prejudice.* New York: Wiley.

Fazio, R. H., Sanbonmatsu, D. M., Powell, M. C., & Kardes, F. C. (1986).

On the automatic activation of attitudes. *Journal of Personality and Social Psychology, 50*, 229–238.

Fiske, S. T. (1998). Stereotyping, prejudice, and discrimination. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *The handbook of social psychology* (Vol. 2 , pp. 357–411). Boston: McGraw-Hill.

Gardner, R. C. (1994). Stereotypes as consensual beliefs. In M. P. Zanna & J. M. Olsen (Eds.), *The Psychology of Prejudice: The Ontario Symposium* (Vol. 7, pp. 1–32). Hillsdale, NJ: Erlbaum.

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

Hager, W., & Hasselhorn, M. (Eds.). (1994). *Handbudh deutschsprachiger Wortnormen* [Handbook of German language word norms]. Göttingen, Germany: Hogrefe.

Haire, M. (1950). Projective techniques in marketing research. *The Journal of Marketing, 14*, 649–656.

Jones, J. M. (1972). *Prejudice and racism.* Reading, MA: Addison-Wesley.

Kahneman, D., Knetsch, J. L., & Thaler, R. H. (1991). Anomalies: The endowment effect, loss aversion and status quo bias. *Journal of Economic Perspectives, 5*, 193–206.

Klineberg, O. (1954). *Social psychology.* New York: Holt, Rinehart and Winston.

Klinger, M. R., & Greenwald, A. G. (1994). Preferences need no inferences? The cognitive basis of unconscious mere exposure effects. In P. M. Niedenthal & S. Kitayama (Eds.), *The heart's eye* (pp. 69–85). San Diego, CA: Academic Press.

Mandler, G. (1982). The structure of value: Accounting for taste. In M. S. Clark & S. T. Fiske (Eds.), *Affect and cognition: The Seventeenth Annual Carnegie Symposium on Cognition* (pp. 3–36). Hillsdale, NJ: Erlbaum.

Medin, D. L., Goldstone, R. L., & Markman, A. B. (1995). Comparison and choice: Relations between similarity processing and decision processing. *Psychonomic Bulletin and Review, 2*, 1–19.

Meyers-Levy, J., & Tybout, A. M. (1989). Schema congruity as a basis for product evaluation. *Journal of Consumer Research, 19*, 39–54.

Murray, H. A. (1943). *Thematic Apperception Test manual.* Cambridge, MA: Harvard University Press.

Newcomb, T. M., Turner, R. H., & Converse, P. E. (1965). *Social psychology: The study of human interaction.* New York: Holt, Rinehart & Winston,

Nosofsky, R. M., & Palmeri, T. J. (1997). An exemplar-based random walk model of speeded classification. *Psychological Review, 104*, 266–300.

Ottaway, S. A., Hayden, D. C., & Oakes, M. A. (2001). Implicit attitudes and racism: The effect of word familiarity and frequency in the Implicit Association Test. *Social Cognition, 19*, 97–144.

Ratcliff, R. (1988). Continuous versus discrete information processing: Modeling accumulation of partial information. *Psychological Review, 95*, 238–255.

Ratcliff, R., & McKoon, G. (1989). Similarity information versus relational information: Differences in the time course of retrieval. *Cognitive Psychology, 21*, 139–155.

Read, S. J., & Miller, L. C. (Eds.). (1998). *Connectionist models of social reasoning and social behavior.* Mahwah, NJ: Erlbaum.

Read, S. J., Vanman, E. J., & Miller, L. C. (1997). Connectionism, parallel constraint satisfaction processes, and Gestalt principles: (Re)introducing cognitive dynamics to social psychology. *Personality and Social Psychology Review, 1*, 26–53.

Reid, P. T., & Holland, N. E. (1997). Prejudice and discrimination: Old paradigms in new models for psychology. In D. F. Halpern & A. E. Voiskounsky (Eds.), *States of mind: American and post-Soviet perspectives on contemporary issues in psychology* (pp. 325–341). New York: Oxford University Press.

Secord, P. F., & Backman, C. W. (1964). *Social psychology.* New York: McGraw-Hill.

Sherif, M., & Sherif, C. W. (1956). *An outline of social psychology.* New York: Harper.

Simpson, G. E., & Yinger, J. M. (1985). *Racial and cultural minorities: An analysis of prejudice and discrimination* (5th ed.). New York: Plenum

Smith, E. R. (1996). What do connectionism and social psychology offer each other? *Journal of Personality and Social Psychology, 70*, 893–912.

Smith, E. R. (1998). Mental representation and memory. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *The handbook of social psychology* (Vol. 1, pp. 391–445). Boston: McGraw-Hill.

Townsend, J. T., & Ashby, F. G. (1983). *The stochastic modeling of elementary psychological processes.* New York: Cambridge University Press.

Zajonc, R. B. (1968). Attitudinal effects of mere exposure. *Journal of Personality and Social Psychology, 9*, 1–27.

Appendix

Words Used in the Experiments

Words Used in Experiment 1

Nonwords: AKAN, BOTSY, BRALDON, CHIG, EKSEN, EMULY, FONNA, FRASK, HOGER, JANK, KAPIE, KROSTIN, LAUMENS, MEDAN, RYIN, SLESHEN, SURA, ZAN

White names: ALAN, BETSY, BRANDON, CHIP, ELLEN, EMILY, DONNA, FRANK, ROGER, JACK, KATIE, KRISTIN, LAUREN, MEGAN, RYAN, STEPHEN, SARA, IAN

Pleasant words: FAMILY, FREEDOM, FRIEND, GENTLE, HAPPY, HEALTH, HEAVEN, HONEST, LAUGHTER, LOVE, LOYAL, LUCKY, PARADISE, PEACE, PLEASURE, RAINBOW, SUNRISE, VACATION

Unpleasant words: ABUSE, ACCIDENT, ASSAULT, CANCER, CRASH, DEATH, DISASTER, DIVORCE, EVIL, HATRED, JAIL, KILL, MURDER, POISON, POLLUTE, POVERTY, SICKNESS, TRAGEDY

Words Used in Experiments 2 and 3

The order in which the nonwords and the translated English words are listed corresponds to the order in which their matching original words are listed. The spelling corresponds to the spelling that participants saw in Experiment 1. In Experiments 2 and 3 all words were spelled according to the identical rules (upper and lower case following regular German spelling).

Insect names (German): AMEISEN, FLIEGE, FLOH, HEUSCHRECKE, HORNISSE, KAKERLAKE, MOSKITO, MOTTE, SCHABE, SPINNE, STECHMÜCKE, TARANTEL, TAUSENDFÜSSLER, TERMITE, WANZE, WESPE

Insect names (English translation): ANTS, FLY, FLEA, LOCUST, HORNET, COCKROACH, MOSQUITO, MOTH, ROACH, SPIDER, GNAT, TARANTULA, CENTIPEDE, TERMITE, BEDBUG, WASP

Nonwords created from German insect names: UMEUKEN, ARIEGE, PLOG, TEPSCHRICKE, LARNIST, HUKERLAUE, MOLKATE, NOSTE, SCHESE, SPILKE, STUCHRASKE, SARANTAT, REUSENDKOSTLER, PERMIFE, NALBER, GIZE

Pleasant words (German): Blumen, Freund, Fröhlichkeit, Geschenk, Geburtstag, Gesundheit, Glück, Humor, Kuβ, Liebe, Musik, Party, Seide, Sommer, Tanzen, Vertrauen

Pleasant words (English translation): flowers, friend, happiness, birthday, gift, health, luck, humor, kiss, love, music, party, silk, summer, dancing, confidence

Unpleasant words (German): Ausrottung, Bedrohung, Beerdigung, Bomben, Gefängnis, Gehässigkeit, Hölle, Krankheit, Mord, Scheidung, Tod, Todeskampf, Totschlag, Verbrechen, Zahnlöcher, Zerstörung

Unpleasant words (English translation): extermination, threat, funeral, bombs, prison, despicableness, hell, illness, murder, divorce, death, death fight, homicide, crime, cavities, destruction.

Received December 30, 1999
Revision received July 5, 2000
Accepted August 5, 2000 ■