A Cognitive Model of the Implicit Associations Test (IAT)

NSF Grant Proposal

Pablo Gomez & Christine Reyna
DePaul University

Goals and Aims

The study of implicit, automatic attitudes has been a dominant theme in social psychology over the last two decades. While we know a great deal about the causes and consequences of such attitudes, we know very little about the specific cognitive processes underlying implicit attitude judgments. We believe that by understanding the cognitive processes subjacent in automatic attitude judgments, our field would be better able to understand the nature of these attitudes. The major aim of the proposed set of studies is to model the cognitive mechanisms underlying one of the most widely used paradigms to test implicit attitudes: the implicit associations test (IAT: Greenwald, McGhee, & Schwartz, 1998).

Since its development six years ago, there have been over 100 articles published using the IAT to test a variety of implicit attitudes. However, despite its popularity, the meaning of IAT scores remains controversial. Little is known about the cognitive processes driving implicit attitudes, and even less is known about what exactly the IAT in particular assesses (Brendl, et al., 2001; De Houwer, J., 2001; Rotherman & Wentura, 2001). Without more research on the cognitive mechanisms underlying the IAT, our field is at risk of creating a literature on implicit attitudes that may be founded on faulty assumptions about the meaning of IAT effects and the underlying cognitive processes they purportedly represent.

By applying a quantitative modeling tool, known as the Diffusion Model (Ratcliff, 1978; Ratcliff & Rouder, 1998; Ratcliff, Gomez & McKoon, 2004)), we hope to shed some light on the exact cognitive processes that are at play when a person takes an IAT. In so doing, we hope to achieve ***three goals***: (1) to advance our field's understanding of the cognitive processes used when participating in the IAT (see theoretical aims); (2) to use the diffusion model to account for IAT data, and to test different versions of the model that are anchored in theories about the locus of IAT effects (see modeling aims); and (3) to establish a data set of different parametric manipulations in the IAT so that future researchers can use the IAT more strategically depending on the cognitive processes they wish to test (see empirical aims).

By effectively modeling the IAT, we would be better able to use this popular methodology more appropriately and effectively. More important, by modeling the cognitive processes at work under various IAT conditions, our field may gain a better understanding of the unique cognitive processes associated with implicit judgments. This later contribution could represent a major theoretical leap in the study of implicit, automatic attitudes.

Background and Significance

Over the last two decades researchers in psychology have generated an important body of empirical and theoretical work on mental processes that occur without conscious awareness (e.g., Bargh, 1989, 1994; Crosby, Bromley, & Saxe, 1980; Devine, 1989; Fazio, Jackson, Dunton, & Williams, 1995; Greenwald & Banaji, 1995; Jacoby & Witherspoon, 1982). In the investigation of attitudes, for example, emphasis has shifted from studying explicit manifestations of attitudes, stereotyping, and prejudice, to investigating more subtle, unconscious, implicit forms (e.g.,

Banaji, 1997; Devine, 2001, Dovidio, Kawakami, Johnson, Johnson, & Howard, 1997; Greenwald & Banaji, 1995; Rudman, Greenwald, Mellott, & Schwartz, 1999; Uhlmann, Dasgupta, Elgueta, Greenwald, & Swanson, 2002).

This relatively new emphasis on automatic, unconscious processes reflects a number of trends in our field. First, social psychologists have renewed interest in automatic processes. Some have claimed that the majority of our daily actions are governed by automatic processes, freeing up resources for more important, complex deliberations (Bargh, 1997; Bargh & Chartrand, 1999). More intriguing, is the growing evidence that even important decisions that should be driven by careful, conscious thoughts can be influenced by attitudes and beliefs in ways that we are not aware of, and some argue, are beyond our immediate control (Bargh & Ferguson, 2000; Bargh, Gollwitzer, Lee-Chai, Barndollar, Troetschel, 2001; Correll, Park, Judd, & Wittenbrink, 2002; Word, Zanna, and Cooper, 1974). Second, this exploding interest in automatic processes has been facilitated by both greater advancements in techniques assumed to measure unconscious or implicit attitudes and their consequences, as well as an awareness that our field's traditional reliance on self-report and other introspective, explicit methodologies has severe shortcomings (Crosby, Gromley & Saxe, 1980; Nisbett & Wilson, 1977).

The Significance of Unconscious and Implicit Processing

Unconscious and implicit cognitive processes can exert a powerful influence on a variety of thoughts and behaviors. For example, we know that our attitudes can be triggered by cues outside of our conscious awareness (Bargh, Chaiken, Govender, & Pratto, 1992; Bornstein, 1989; Eagly, Ashmore, Makhijani, Longo, 1991; Fazio, Sanbonmatsu, Powell, & Kardes, 1986, Zajonc, 1968). Even socially undesirable attitudes, like prejudice, can be invoked without our conscious control by incidental exposure to stimuli or contexts that we associate with the group in question (Banaji, Hardin, & Rothman, 1993; Devine, 1989; Dovidio, Evans, & Tyler, 1986, Gilbert & Hixon, 1991; Gaertner & McLaughlin, 1983; Perdue & Gurtman, 1990; Paulhus, Martin, & Murphy, 1992). More important, these implicit attitudes can manifest themselves in behaviors that can have damaging social consequences. For example, unconscious prejudice towards ethnic minorities can influence job or academic selection decisions, can undermine interpersonal interactions, and can compromise our cognitive resources and efficiency on interdependent tasks with members of minority groups (e.g., Dovidio, Gaertner, Kawakami, & Hodson, 2002; Dovidio, Kawakami, & Gaertner, 2002; Richeson & Shelton, 2003). Implicit prejudice can also have potential life-threatening consequences. For example, in a police simulation paradigm, automatic stereotyping has been associated with facilitated reactions to shoot Black criminals faster than White criminals, as well as increased errors resulting in decisions to shoot Black, unarmed citizens compared to White, unarmed citizens (Correll, Park, Judd, & Wittenbrink, 2002).

The Implicit Associations Test (IAT)

To some degree, advancements in the study of unconscious processes have been limited by our ability to effectively measure them. A number of methodologies have been developed to assess unconscious processes including word fragment completion (Bassili & Smith, 1986; Gilbert & Hixon, 1991; Hetts, Sakuma, & Pelham, 1999), semantic and subliminal priming (Dijksterhuis, Aarts, Bargh, & van Knippenberg, 2000; Dovidio, Kawakami, Johnsom, Johnson, & Howard, 1997; Fazio, Jackson, Dunton, & Williams, 1995; Greenwald, Klinger, & Liu, 1989; Higgins, 1989; Krosnick, Betz, Jussim, & Lynn, 1992; Roediger, 1990; Roediger & Blaxton,

1987), and even physiological assessments (e.g., Blascovich, Mendes, Hunter, Lickel, Kowai-Bell, 2001; Vanman, Paul, Ito, Miller, 1997). However, no method has gained as much prominence over the last few years as the Implicit Associations Test (IAT: Greenwald, McGhee, & Schwartz, 1998). Since its introduction six years ago, over 100 articles using this methodology have been published by scholars all over the globe. The IAT is now the most popular, and perhaps the most controversial, methodology for measuring implicit attitudes (Gawronski, 2002).

As its name implies, the IAT is assumed to measure the associative strength between two target concepts or categories (e.g., Black vs. White) and positive or negative attributes (e.g., pleasant vs. unpleasant). This test has been used to measure implicit associations or attitudes by comparing the response latencies when each category is paired with positive versus negative words. Using a race IAT as an example, participants are exposed to images of African-American or Caucasian faces and positive or negative words (each face is followed by a word). They are asked to make the same response (e.g., pressing a computer key with the right hand) when the target image is of an African-American OR the word is positive. They are asked to make a different response (e.g., pressing a key with the left hand) when the image is Caucasian OR the word is negative. If participants have positive attitudes toward African-Americans, then this task is relatively simple because they can combine the two categories (African-American/positive) into a single unit: things that are positive. This should facilitate faster responses to this task. However, if participants have negative attitudes towards African-Americans, then they would have to make an affective adjustment after categorizing a face as African-American (a negative attitude object), and then categorizing a word as positive using the same response protocol. This should produce a delay in response latency because the same response protocol (e.g., right-hand key press) is required for two attitudinally diametric judgments. The same phenomena occur when pairing Caucasian faces with either positive or negative words. Thus, one can compare the response latencies of all of the face/word combinations to see which attitudes were best facilitated and which were inhibited. These response latencies are often interpreted as a measure of implicit attitudes towards African-Americans and Caucasian-Americans.

The IAT has been used to measure a wide range of attitudes, including racial prejudice (Greenwald, McGhee, & Schwartz, 1998; Rudman, Greenwald, Mellott, & Schwartz, 1999), attitudes towards obesity and implicit weight identity (Grover, Keel, & Mitchell, 2003; Teachman, Gapinski, Brownell, Rawlins, & Jeyaram, 2003), attitudes toward the elderly (Jelenec, & Steffens, 2002), and attitudes towards gay men (Steffens & Buchner, 2003). It has also been used to study ingroup bias (Rudman, Feinberg, & Fairchild, 2002), self-esteem (Greenwald & Farnham, 2000), and gender identity (Aidman, & Carroll, 2003). In addition, the IAT has been used to examine attitudes toward behaviors and intentions to engage in behaviors, such as cigarette smoking (Chassin, Presson, Rose, Sherman, & Prost, 2002; Sherman, Rose, Koch, Presson, & Chassin, 2003) and drinking alcohol (Palfai & Ostafin, 2003). Finally, the IAT has been used to investigate psychological disorders such as anxiety (Egloff & Schmukle, 2002), phobias (Teachman, Gregg, & Woody, 2001), and reactions to violence among psychopathic murderers (Gray, MacCulloch, Smith, Morris, & Snowden, 2003). Thus, the IAT has become a powerful tool used to study a wide variety of issues that are of central importance to psychologists and social scientists.

Problems to be Addressed: Theoretical and Psychometric Limitations of the IAT

Despite the IAT's popularity as an attitude assessment tool, the methodology is not without controversy (Brendl, Markman, & Messner, 2001; Gawonski, 2002, also see Fazio & Olson, 2003). The assumption made by researchers who use the IAT is that the results of the IAT

reflect implicit attitudes; however, many questions remain about the cognitive processes driving implicit attitudes, and there is even more uncertainty about what exactly the IAT in particular assesses (Brendl, et al., 2001; De Houwer, J., 2001; Mierke & Klauer, 2001; Olsen & Fazio, 2004; Rotherman & Wentura, 2001). For example, concern has arisen over the fact that many measures of implicit or unconscious attitudes do not correlate well with each other (Bosson, et al., 2000; Brauer, M., Wasel, W., Niedenthal, P., 2000; Sherman, Presson, Chassin, Rose, & Koch, 2002; however, see Cunningham, Preacher, & Banaji, 2001). This suggests that the different paradigms tap into different psychological processes or constructs. Also, there is conflicting support for the test's predictive validity. Some scientist have found that the IAT does correlate with psychological and behavioral variables associated with the target attitudes under investigation (Greenwald & Furnham, 2000; Jordan, Spencer, & Zanna, 2002; McConnell & Liebold, 2001), while others have found little or no association (Karpinski & Hilton, 2001). IAT findings also seem to differ depending on method variance, such as how the targets are conceptualized (Dasgupta & Greenwald, 2001; Mitchell, et al., 2001; Wittenbrink, Judd, & Park, 2001), and the experimental context (Lowery, et al., 2001, Richeson & Ambady, 2003). The test also shows some sensitivity to motivation (Lowery, et al., 2001), which calls into question whether or not the test is truly tapping into implicit processes.

These findings reflect a larger uncertainty about the cognitive mechanisms underlying the IAT and other implicit attitude measures. Recently, some scientists have called into question whether this popular paradigm is actually measuring attitudes at all, and several alternatives have been proposed to explain the underlying cognitive processes at work in the IAT. For example, Rothermund & Wentura (2001) suggest that IAT effects cannot be accounted for by evaluative association alone, and that figure-ground asymmetries also seem to be contributing to this effect. They found that results from the IAT can be altered depending on which of the two target categories is considered "figure" and which is considered "ground." Mierke & Klauer (2001) proposed that IAT effects are due to differential costs in task-switching depending on the IAT condition. Specifically, they found that switching between tasks was associated with significantly more "costs" in the incompatible IAT phase; however, when task-set reconfiguration was possible in advance, the magnitude of the IAT effect was reduced. While these studies point to some variance accounted for by features of the IAT task or method, they cannot fully account for IAT effects. In addition, neither of these alternative perspectives provides adequate modeling of the cognitive mechanisms at work in the IAT. A more promising approach comes from research by Brendl and colleagues (Brendl, et al., 2001), who did attempt to apply a computational model to account for the IAT effects. They found that people adjust their response criteria with the difficulty of the combined tasks. Criteria adjustment can be a function of a number of things other than associative strength, and thus the meaning of the IAT remains obscure.

Intellectual Merit and Broader Impact of the Proposed Research

At this point in time, our field has to contend with a methodology that is widely used but poorly understood. Without more research on the cognitive mechanisms underlying IAT effects, our field could face a crisis. With so many researchers using this paradigm, we are creating a burgeoning literature on implicit attitudes that may be founded on faulty assumptions about the meaning of IAT effects and the cognitive processes they purportedly represent. By effectively modeling the cognitive processes at work in the IAT, we would be better able to use this popular methodology more appropriately and effectively.

Through the proposed research, we hope to achieve three goals: (1) to advance our field's understanding of the cognitive processes used when participating in the IAT (see theoretical aims); (2) to use the diffusion model to account for IAT data, and to test different versions of the model that are anchored in theories about the locus of IAT effects (see modeling aims); and (3) to establish a data set of different parametric manipulations in the IAT so that future researchers can use the IAT more strategically depending on the cognitive processes they wish to test (see empirical aims).

In the long term, by modeling the cognitive processes at work under various IAT conditions, our field may gain a better understanding of the unique cognitive processes associated with implicit attitudes. This later contribution could represent a major theoretical leap in the study of implicit, automatic attitudes. In addition, by understanding these underlying processes, scientists and clinicians will be in a better position to use research on implicit attitudes to develop interventions for dysfunctional associations such as phobias, low self-esteem, and prejudice.

Theoretical Aims

Many researchers use the IAT as a tool to measure attitude associations, and are less concerned about the underlying mechanisms. They assume that the IAT score represents the strength of association between two categories, and are less specific about the illusive cognitive mechanisms that produce the score. With the exception of isolated attempts to model the IAT, most social psychologists that use this paradigm apply basic assumptions about the cognitive processes associated with automatic thought to IAT effects. The dominant assumption regarding the IAT is that constructs that are highly associated in memory will be easier to pair together into a single response (a key press). Therefore, cuing one category (e.g., "African-American") will facilitate responses in a judgment task if the response is attitudinally compatible, and inhibit responses that are incompatible (e.g., Banaji, Lemm, & Carpenter, 2004). In other words, stimuli are easier to categorize in the attitudinally congruent conditions of the IAT than in the incongruent conditions.

This description of automatic attitude activation draws from models of semantic memory in which information is thought to be stored in memory in a type of network, such that concepts that are highly associated are linked more directly than concepts that are weakly associated. When one piece of information is cued in memory, its close associates are also activated and thus are more readily brought to mind with the cueing of the original piece of information—commonly referred to as "spreading activation" (Collins & Loftus, 1975). The processing assumption that this kind of theory makes about IAT scores is that participants are better able to discriminate between target cues when the response mappings (which response goes with which decision) are congruent with pre-existing associations with the social category.

Alternative Perspectives.

Ideas inspired in spreading activation formulations have dominated thinking about the cognitive processes that underlie automatic attitudes; however, two important theoretical issues must be considered. First, in the cognitive literature, the notion of spreading activation has been strongly challenged. Other theories claim that activation does not spread; instead, concepts are highly related because they co-occur often, and form a compound-cue in memory (cf. Ratcliff & McKoon, 1995). For example, the words insect and dirty might co-occur in discourse more often than the words insect and love (see Karpinski & Hilton, 2001and Olson & Fazio, 2004 for a related perspective involving the IAT). It should be noted that the spreading activation notion has been used more as a heuristic to understand implicit attitudes than as a computational model of implicit processes. This leads us to the second issue, and certainly the most important of the two. Because spreading activation has been used more as a heuristic to talk about implicit associations than as a real model of what goes on in the cognitive system, thinking about the IAT in these terms may have hindered the development of more sophisticated models.

In order to map the cognitive processes underlying the IAT, we turn away from conventional spreading activation based intuitions and employ a model that makes explicit assumptions about processing, and explicit quantitative predictions: the diffusion model (Ratcliff, 1978; Ratcliff & Rouder, 1998; Ratcliff & Smith, 2004). This is a model of decision making in tasks in which participants make fast choices (like the IAT). The basic intuition behind the model is that decisions are made after evidence is accumulated over time. While diffusion models have revolutionized the cognitive literature on decision processes[1], they have yet to be applied in social cognitive contexts (however, see Brendl, et al., 2001 for a discussion of a random walk model which is similar in some respects to a diffusion model).

Diffusion models reveal a number of interesting alternative explanations for decision processes that could illuminate our understanding of tasks that intend to measure implicit attitudes. We know that cuing a category like race can produce a number of different information processing strategies; however, with diffusion modeling, we now have the tools to measure which strategies are being used in which contexts (such as in the IAT). For example, cuing a category may lower the decision threshold for category-consistent information and raise it for category-inconsistent information. We already know that stereotype-consistent information seems to be processed more readily and that perceivers "pay more attention" with information that violates the stereotype. Diffusion models allow us to map this process precisely. Another possibility is that perceivers may be just as proficient at distinguishing positive and negative information following a category cue, but they hesitate at the response execution stage. For example, a person with a spider phobia may recognize a word as positive just as fast as recognizing a word as negative after seeing an image of a spider, but they hesitate when it comes to delivering a response that would confirm this evaluation. There are other strategies that can be mapped using diffusion models that will be described in more detail below.

---

[1] In the last few years, the diffusion model has been able to account for data in a wide variety of experimental paradigms in which participants choose between two possible responses (Busemeyer & Townsend, 1993; Ratcliff & Rouder, 1998; Ratcliff, Gomez, McKoon, 2004; Roe, Busemeyer, & Townsend, 2001; Smith, 1995).

Modeling Aims

Our goal is to model competing hypotheses about the locus of the IAT within an explicit computational framework. To do so, we need a model in which we can translate hypotheses about the task into assumptions about the behavior of the different parameters of the model (where each parameter corresponds to different components of the cognitive process, i.e., the parameters of the model would have psychological meaning). The diffusion model fits this requirement. In the diffusion model, we can allow different parameters to change across conditions (cf. Gomez, Perea, Ratcliff & McKoon, 2004) in order to represent hypotheses about what processes change across such conditions. We can then select the most efficient model by comparing the quality of their fits (and penalizing for the number of free parameters). This way, we will be able to shed light on the cognitive process involved in the IAT. In addition, this model fits not only the mean response time (RT), it fits the response probabilities and the RTs for correct and error responses.

Description of the model. Figure 1-A in the Appendix shows a representation of a trial according to the diffusion model. The assumption is that the response time in a dual choice task is the sum of three processing components. The first component is the encoding process. When a stimulus is presented (e.g., a face or a word), the participant extracts the physical and psychological features that are relevant for the discrimination task at hand (e.g., the color of the skin, the shape of the nose, or the morphemes of the word). This process feeds a strength-of-activation value into the second component: the decision process. The model assumes that the information relevant to the discrimination task (e.g., race of the target, valence of a word in the IAT) is accumulated over time toward one of two decision boundaries (e.g., whether the face is African-American or White, whether the word is positive or negative). The accumulation of evidence is noisy (hence the jagged line), meaning that within a trial the rate of accumulation of evidence varies around a mean value (called the drift rate). We refer to this noisy accumulation of evidence as the "diffusion process." The distances from the starting point of the accumulation process to each of the boundaries can be thought of as the amount of evidence necessary to make a decision. When the diffusion process reaches one of the decision boundaries, the decision is made and the response is initiated. The response execution stage is the third component depicted in Figure 1, and it is assumed to be independent of the other two components. The diffusion model is very good at estimating the different components associated with the decision process; however, the time taken by the encoding and response processes is grouped into a single parameter in the model. For a more detailed explanation of the parameters of the model go to the Appendix.

The competing hypotheses. In order to identify the cognitive mechanisms underlying the IAT, we will use the parameters of the diffusion model to test which of the four hypotheses below best fits the data. Rather than predicting IAT effects to be the result of a single mechanism, we recognize that implicit processes may be the result of multiple strategies that are employed under different circumstances. Therefore, we predict that a number of the proposed strategies could be at play during the IAT depending on the conditions surrounding the IAT. Therefore, we have designed our set of studies to test all four hypotheses. To this end, we will hold different parameters constant across different experimental conditions while allowing other parameters to vary.

Increased Discriminability Hypothesis. Perhaps the most traditional interpretation of the typical IAT results is that, due to the pre-existing associations between categories and valences, subjects' ability to discriminate the stimuli improves in the congruent condition relative to the incongruent condition. This notion is implemented in the first model under consideration, which assumes that the subjects' ability to discriminate along the relevant dimensions in the IAT is affected by the response mapping. In the model, this is captured by the average rate of accumulation of information (*drift rate* parameter). For this reason, the first model will have the *drift rates* as free parameters (see Panel A in Figure 2).

Decision Criteria Hypothesis. Another interpretation of the IAT results (cf. Brendl et al, 2000), is that the different IAT conditions might lead subjects to change their decision thresholds (i.e., subjects might prioritize speed versus accuracy, or might require different amounts of evidence to make a categorization task). In the model, this is implemented by assuming that participants might change their decision criteria across blocks or conditions in the IAT. In the model this hypothesis is captured by allowing the parameters related to location of decision

boundaries to change across task. These parameters are *a* and *z*. (see Panel B in Figure 2.)

Bias in the Decision Process Hypothesis. Another possible mechanism underlying the IAT is that subjects might perceive the stimulus in a biased manner[2]. For example, in the incongruent condition all words (even the ones nominally "positive") might seem more negative than in the congruent condition. Note that according to this hypothesis, the ability of subjects to discriminate stimuli does not change; instead, the extraction of information from the stimulus is biased towards one of the two alternatives. The parameters of the model that need to change to implement this hypothesis are the *drift rates.* All drift rates, however, would change by a constant added or subtracted to them.

Encoding and Response Execution Hypothesis. The final mechanism under consideration is not related to the decision process per se. The assumption is that the IAT effects are a consequence of the response execution components of the RT, the idea being that in the incongruent condition the response procedure is more complex than in the congruent condition (see Mierke & Kaluer, 2001 for a compatible formulation). In other words, according to this hypothesis, subjects' ability to discriminated between the valences and categories in the IAT would not change, neither would their criteria to response; instead, subjects would tend to have difficulty remembering which key corresponds to which decision in the incongruent condition. The parameter related to the response execution process is T_{er}, so, to implement this hypothesis T_{er} would be free to vary across IAT conditions.

Overdetermination and Model Selection. Scholars of socio-cultural issues have, for a long time, recognized that phenomena are over-determined, meaning that there are many causes behind an event. We expect that IAT scores could also be a consequence of many cognitive mechanisms. An advantage of our modeling strategy is that the hypotheses described above need not be mutually exclusive. We can implement hybrid models, where we assume that the IAT effects are related to more than one process. The procedure that we will use to compare and test all the implemented models is based on the notion that there is a tradeoff between parsimony and quality of fits to data. Specifically, we will use model selection methods, such as BIC (Schwarz, 1978) or PBCM (Wagenmakers, Ratcliff, Gomez & Iverson, 2004) to select the most efficient model for each dataset (i.e., all hypothesized models will be fitted to the data from each experiment).

Empirical Aims: Research Design and Methods

We have two fundamental goals in the proposed series of experiments: first, we want to collect IAT data across a variety of domains; second, the data collected has to provide enough constraints to the hypotheses being tested. In modeling quantitative models like the diffusion model, it is essential to perform parametric manipulations on variables relevant to the task. For example, in the IAT we might use words that have been rated as "very positive", "very negative", "slightly positive", or "slightly negative". Below, we present some of the details about the proposed studies.

---

[2] We use the word "bias" in terms of a cognitive process rather than as a social construct. Here we mean that the subjects might be more likely to interpret the stimulus as a member of one category than of the other.

Participants

Forty undergraduates from DePaul University will be recruited for each of the seven studies presented below (totaling 280 participants). Special attention will be made to recruit women and people of color.

Materials

Each of the seven studies will be conducted on individual computers. Inquisit software will be used to program the IAT studies into the computers and to record participants' data. Digital cameras will be used to take photos of the target category exemplars (faces of African-American and European-American males and females). To be able to constrain the diffusion model, it is necessary to perform parametric manipulations, and to collect enough data per condition. The parametric manipulations that we are proposing across all experiments relate to including stimuli across different levels of the valence dimension (e.g, words that are highly positive, words that are less positive, word that are highly negative, and words that are less negative; e.g. Siegle, 1994), and also, stimuli that vary across the category exemplars (e.g., Caucasian vs. African American features; see Maddox, in press for a example of the effects of phenotypical features).

Procedures

Each of the seven studies will be based on a standard IAT procedure as specified in Greenwald, et al., 1998. In most of the proposed studies, modifications will be made in either the procedures or the stimuli in order to test different possible cognitive processes (see below), and to collect more data from each subject. In particular, our experiments will include more blocks than the standard IAT described below. Our goal is to collect twice as much data from each participant than in the typical IAT study, therefore, our participants will go through the five blocks described below twice.

Standard IAT procedure. In the standard IAT procedure, participants are instructed to focus their attention on a computer screen. They are told that they will see either words or images appear in the center of the screen. On both the left and right side of the word or image will be category labels. Participants will be asked to indicate, as quickly as possible, which category the word or image belongs to. So, for example, if the category on the left of the screen is "African-American" and the category on the right of the screen is "European-American," participants will have to press a left hand key when the image is of an African-American and a right hand key when the image is of a European-American. They will classify words as either positive or negative using a similar procedure. After participants become acquainted with the categories and the keys associated with them, they will have to combine the categories into the same key presses, such that they would press a right hand key when the image is of an African-American OR the word is positive, and a left-hand key when the image is of a European-American OR the word is negative.

In the typical IAT study, participants complete five steps associated with the IAT. The first step is the *initial target discrimination task*, in which participants will be trained to identify and respond to faces that are either African-American or European-American by pressing a left or right-hand key that is assigned to each category (e.g, pressing the "D" key with the left hand for African-American faces, and the "K" key with the right hand for European-American faces). The second step is the *associated attribute discrimination task*. This task is identical to the target discrimination task, but this time, participants will be asked to discriminate between positive and

negative words by pressing the "D" key when a word is positive and the "K" key when the word is negative. The third step is the *initial combined task*. Here participants will be presented with both images and words and will have to classify each one using the same key if the image is of an African-American OR the word is positive (the D key), and will use a different key (the K key) if the image is of a European-American OR the word is negative. The fourth step, the *reversed target concept discrimination task*, retrains the participants to switch the hands they associate with the different race categories, such that now they would press the D key if they see an image of a European-American and the K key if they see an image of an African-American. Positive and negative words remain associated with their initial key assignments. The fifth task is the *reversed combined task*, wherein participants now must categorize European-American images OR positive words using the D key and African-American images OR negative words using the K key. Response times associated with how fast participants can categorize images and words in the combined tasks will be recorded. Difference scores using these response times for each category-word type pairing will be calculated and used to measure facilitation or inhibition in the combined tasks.

Study 1: We first want to conduct a preliminary analysis of a race-based IAT to see which of the different diffusion model hypotheses best fit with the obtained data (see also Pilot Study below). We are interested in determining if the same cognitive processes are employed when a race category is cued using race-specific names (the pilot study) versus race-specific images. Therefore, Study 1 will model a standard IAT procedure using faces and the results will be compared to the best fitting model using names in the IAT (the pilot study). To summarize, there will be three within subject variables manipulated: block in the IAT (five blocks); valence if the words (four levels); typicality of the face (i.e., four levels of Caucasian vs. African-American facial features).

Study 2: In this study, we will manipulate the discriminability of the stimuli. We will employ a standard IAT procedure using images and positive and negative words. However, to test the idea that paired constructs make discriminating stimuli more effective, we will "corrupt" the stimuli to make perceiving them more challenging. To do this, we will partially mask the stimuli so that they are discernable but difficult to perceive without effort or some type of cognitive facilitation. To summarize, there will be four within subject variables: block in the IAT (five blocks); valence of the words (four levels); typicality of the face (i.e., four levels of Caucasian vs. African-American facial features); and SOA between the target stimuli and the mask (three levels).

Study 3: Studies 3 and 4 are designed to manipulate the response execution processes. Shorter response times would be generated by a hesitation that occurs at the response stage and not on stages that involve perceiving the stimuli. In Study 3, we will modify the standard IAT by switching hands (key presses) for all combinations of category variables. The standard IAT procedure only switches hands for classifying race. We will fully counterbalance stimulus category with which key is pressed. In doing so, we will be able to determine the contribution of altering the response mapping for the categorization task to the standard IAT results. In sum, there will be three within subject variables manipulated: block in the IAT (ten blocks); valence if the words (four levels); typicality of the face (i.e., four levels of Caucasian vs. African-American facial features).

Study 4: Study 4 will test the response execution model by employing a "go/no-go" modification in the IAT instructions. Gomez et al (2004) have shown that the go/no-go procedure affects the response execution stage. Participants will be told to press the D key if they

see an African-American face OR a positive word, and do nothing if they see European-American faces OR negative words and just wait for the next trial. In sum, there will be four within subject variables manipulated: block in the IAT (five blocks); valence if the words (four levels); typicality of the face (i.e., four levels of Caucasian vs. African-American facial features); and go/no-go instructions (e.g, "go" to positive words, or "go" to negative words).

Studies 5 and 6: In studies 5 and 6 we will manipulate the decision criteria. In Study 5, we will manipulate speed/accuracy instructions to test whether participants will increase or decrease their decision threshold. Speed instructions should make lowering the threshold more likely while accuracy instruction could raise the threshold. There will be four within subject variables manipulated: block in the IAT (five blocks); valence if the words (four levels); typicality of the face (i.e., four levels of Caucasian vs. African-American facial features); and speed vs. accuracy instructions.

Study 6 will attempt to achieve the same effects for each decision separately by manipulating motivation to respond in one way versus the other. We will vary the "payoffs" (points assigned to certain stimuli) for key presses. For example we might instruct participants that they will get 10 points for correct responses to negative items and 20 point for correct responses to positive items. This way, we will be able to change the decision threshold for each category and valence separately. There will be four within subject variables manipulated: block in the IAT (five blocks); valence if the words (four levels); typicality of the face (i.e., four levels of Caucasian vs. African-American facial features); and payoff levels (two levels).

Study 7: In Study 7, we will vary the base rates of the different categories that are presented (vary the number of category exemplars participants are exposed to). In so doing, we will be biasing the accumulation of evidence process. There will be four within subject variables manipulated: block in the IAT (five blocks); valence of the words (four levels); typicality of the face (i.e., four levels of Caucasian vs. African-American facial features); and base rates (two levels).

Each of these studies is designed to manipulate different possible cognitive process at work in the IAT. While we intend to explore the parameters under which different cognitive strategies (the different models) might be at play in different testing contexts, it is important to note that any of the possible models could apply to any of the seven studies.

Model selection procedures. We anticipate using two model selection techniques: information criteria and montecarlo bootstrapping. A combination of the two model selection techniques will provide us with complementary information. The first one is based on criteria that trade off descriptive accuracy (i.e., goodness of fit) of a model against model complexity as measured by the number parameters in the models (Akaike, 1973, 1974; Schwarz, 1978). The two most widely used information criteria of this type are the Akaike Information Criterion (AIC), and the Bayesian Information Criterion (BIC). The formula for the AIC is

$$AIC_i = -2\ln L_i + 2K_i,$$

while the formula for the BIC has a very similar form:

$$BIC_i = -2\ln L_i + \log(N)K_i.$$

In the two formulas, $\ln L_i$ is the natural logarithm of the maximum likelihood and K_i is the number of parameters in model i. In the BIC, the penalty term depends not only on the number of free parameters, but also on the number of observations N.

The other method that we plan to use is based on montecarlo simulations and it is particularly useful when comparing two models (Parametric Bootstrap Cross-fitting Method or PBCM; Wagenmakers, Ratcliff, Gomez & Iverson, 2004). This method determines how

diagnostic a difference in goodness of fit might be based on the degree to which the models mimic each other. The basic idea is that data is generated from one of two competing models, and then both models are fit to the simulated data. The difference in goodness of fit is calculated and when this procedure is repeated hundreds of times, two distributions of differences in goodness of fits arise: one for when the first model generates the data, and the other for when the second model generates the data. The point at which the two distributions intersect can be considered the optimal decision criterion for model selection.

Appendix

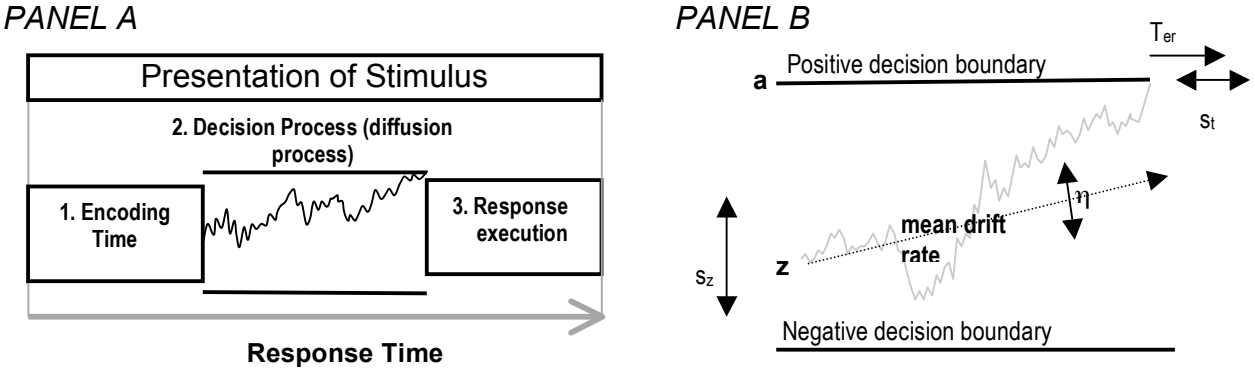

Figure 1. The figure shows a representation of the sequence of events in a trial of a dual-choice task in which the stimulus is presented until a response is made. See text for details. Panel B shows An illustration of the diffusion model. The starting point of the accumulation process is labeled z and the upper boundaries are labeled a. The rate of accumulation of information is called the drift rate, and it is determined by the quality of the information extracted from the stimulus. For example, in a task where faces are classified as African-American or Caucasian, faces with prototypical Caucasian features would have a large mean value of the drift rate toward the "Caucasian" boundary. There is noise (variability) in the process of accumulating information so that processes with the same mean drift rate do not always terminate at the same time, thus, producing reaction time (RT) distributions. In addition, they do not always terminate at the same boundary, thus producing errors. This variability is called "within trial" variability. Panel B in Figure 1 shows one process with the mean drift rate represented by the arrow and the accumulation of noisy information represented by the jagged line.

Components of processing are assumed to be variable across trials. Drift rate is assumed to be normally distributed with standard deviation η and starting point is assumed to be uniformly distributed with range sz. There are nondecision components of processing such as encoding and response execution that are not part of the decision process. These are combined in the diffusion model into one component with mean T_{er} . The nondecision component of processing is assumed to have variability across trials and it is assumed to be uniformly distributed with range s_t. In sum, the parameters of the diffusion model correspond to the components of the decision process as follows: z is the starting point of the accumulation of evidence. a is the upper boundary, the lower boundary is set to 0, η is the standard deviation in mean drift rate across trials, sz is the range of the starting point across trials, and s_t is the range of T_{er} across trials. For each different stimulus condition in an experiment, it is assumed that the rate of accumulation of evidence is different and so each has a different value of drift, v. Within-trial variability in drift rate (s) is a scaling parameter for the diffusion process (i.e., if it were doubled, other parameters could be multiplied or divided by two to produce exactly the same fits of the model to data).

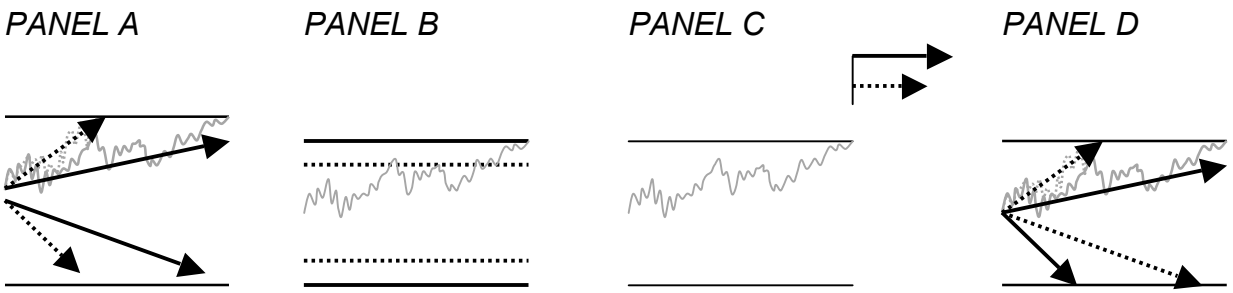

Figure 2. The figure shows representations of the four hypotheses about the locus of the IAT effects. In all Panels the thick solid line represents the incongruent condition, and the thick dotted line represents the congruent condition in the IAT. Panel A represents the hypothesis that discriminability increases in the congruent condition, and hence, the accumulation of information is faster towards the correct decision boundaries. Panel B represents the hypothesis that the decision criteria change across conditions, and hence, the decision boundaries are closer to the starting point in the congruent condition than in the incongruent condition. Panel C represents the hypothesis that the nondecision components (e.g., response execution) change across conditions. Panel D represents the hypothesis that there is a bias in the accumulation of information process.