Nosek, B. A., Greenwald, A. G., & Banaji, M. R. (in press). The Implicit Association Test at age 7: A methodological and conceptual review. In J. A. Bargh (Ed.), Automatic processes in social thinking and behavior. Psychology Press.

The Implicit Association Test at Age 7: A Methodological and Conceptual Review

Brian A. Nosek

University of Virginia

Anthony G. Greenwald

University of Washington

Mahzarin R. Banaji

Harvard University

Authors' note: This research was supported by the National Institute of Mental Health, MH-41328, MH-01533, MH-57672, and MH-68447 and the National Science Foundation, SBR-9422241 and SBR-9709924. Correspondence concerning this article should be addressed to Brian Nosek, Department of Psychology, University of Virginia, Box 400400, Charlottesville, VA, USA, 22911; or to nosek@virginia.edu.

Among earthly organisms, humans have a unique propensity to introspect or look inward into the contents of their own minds, and to share those observations with others. With the ability to introspect comes the palpable feeling of "knowing," of being objective or certain, of being mentally in control of one's thoughts, aware of the causes of one's thoughts, feelings and actions, and of making decisions deliberately and rationally. Among the noteworthy discoveries of 20th century psychology was a challenge posed to this assumption of rationality. From the groundbreaking theorizing of Herbert Simon (1955) and the mind boggling problems posed by Kahneman and Tversky (1974) to striking demonstrations of illusions of control (Wegner, 2002), the paucity of introspection (Nisbett and Wilson, 1977), and the automaticity of everyday thought (Bargh, 1997), psychologists have shown the frailties of the minds of their species.

As psychologists have come to grips with the limits of the mind, there has been an increased interest in measuring aspects of thinking and feeling that may not be easily accessed or available to consciousness. Innovations in measurement have been undertaken with the purpose of bringing under scrutiny new forms of cognition and emotion that were previously undiscovered and especially, by asking if traditional concepts such as attitude and preference, belief and stereotype, self-concept and self-esteem can be rethought based on what the new measures reveal.

These newer measures do not require introspection on the part of the subject. For many constructs this is considered a valuable, if not essential, feature of measurement; for others, avoiding introspection is greeted with suspicion and skepticism. For example, one approach to measuring math ability would be to ask "how good are you at math?" whereas an alternative approach is to infer math ability via a performance on a math skills test. The former requires introspection to assess the relevant construct, the latter does not. And yet, the latter is accepted as a measure of math ability, and is even preferred to one requiring self-assessment.

When the target construct concerns a preference, stereotype, or identity rather than performance, issues about interpretation turn out to be more complex those involving performance such as on tests of ability (memory, concept formation) where there is an assumed correct answer. For some, the dismissal of introspection as relevant to the assessment of such constructs is difficult. Attitudes, stereotypes, and identity appear to be wound so tightly to subjective thoughts and feelings, that there "asking" seems to be the most persuasive of probes. "I know how I feel" and "Don't tell me how I feel" are not just expressions in interpersonal communication – they are assumed by psychologists who accept such an epistemological stance.

In the last few years, one measure in particular, the Implicit Association Test (IAT), has spurred discussion among both experts and non-experts – about its mechanisms, scope, interpretation, and political implications. In this chapter, we review the main issues that are debated and provide our best assessment of its current status.

Implicit cognition. Building on the implicit-explicit distinction in memory (Roediger, 1990; Schacter, Bowers, and Booker, 1989), Greenwald and Banaji (1995) proposed a general distinction for implicit cognition. They defined an implicit construct as "the introspectively unidentified (or inaccurately identified) trace of past experience that mediates R" where R refers to the category of responses that are assumed to be influenced by that construct (Greenwald & Banaji, 1995, p. 5). Greenwald and Banaji applied that general definition to social psychology's most central constructs – attitudes, stereotypes, and self-esteem. They noted that implicit cognition could reveal associative information that people were either unwilling or unable to report. In other words, implicit cognition could reveal traces of past experience that people might explicitly reject because it conflicts with values or beliefs, or might avoid revealing because the expression could have negative social consequences. Even more likely, implicit cognition can reveal information that is not available to introspective access even if people were motivated to retrieve and express it (see Wilson, Lindsey, and Schooler, 2000 for a similar theoretical distinction for the attitude construct specifically). Such information is simply unreachable in the same way that memories are sometimes unreachable, not just in amnesic patients but in every person.

For many constructs such as memory, attitudes, stereotypes, self-concept, self-esteem, personality, and knowledge, the implicit-explicit taxonomy has not just helped to organize existing theory and empirical evidence, but has also broadened the construct beyond introspective limits. For example, while few definitions of attitude mentioned introspective access as a necessary feature, until the 1980s attitude measurement largely proceeded as if the very definition of attitude relied on an assumption that attitudes were consciously available (Greenwald & Banaji, 1995).

Implicit measurement. Whatever the value of the implicit-explicit distinction, in practice the distinction has been rather loosely applied to organize a heterogeneous set of assumed cognitive mechanisms. The term *implicit* has come to be applied to measurement methods that (a) avoid requiring introspective access, (b) decrease the mental control available to produce the response, (c) reduce the role of conscious intention, and (d) reduce the role of self-reflective, deliberative processes. The next generation of research in implicit cognition will likely revise the simple implicit-explicit distinction and introduce a more refined taxonomy that better reflects the heterogeneity of cognitive processes that are collectively termed *implicit*. In this chapter, we do not tackle these issues. Instead, we spotlight a particular method and summarize the evidence for its reliability, validity, interpretation, and proper use.

The Implicit Association Test (IAT). The focus of this review is on the Implicit Association Test (IAT; Greenwald, McGhee, & Schwartz, 1998). In the seven years since its initial publication, the IAT has been applied in a diverse array of disciplines including social and cognitive psychology (Greenwald & Nosek, 2001; Fazio & Olson, 2003), clinical psychology (de Jong, Pasman, Kindt, & van den Hout, 2001; Teachman, Gregg, & Woody, 2001), developmental psychology (Baron & Banaji, 2004; Dunham, Baron, & Banaji, 2004), neuroscience (Cunningham, Johnson, Raye, Gatenby, Gore, & Banaji, in press; Phelps et al., 2000; Richeson et al., 2003), market research (Maison, Greenwald, & Bruin, 2001), and health psychology (Teachman, Gapinski, Brownell, Rawlins, & Jeyaram, S., 2003). The wide range of application provides an ample research literature from which to review the features, strengths and limitations of the IAT for continued research of implicit cognition. The present review can be seen as an "age 7" follow-up to Greenwald and Nosek's (2001) review of the IAT "at age 3" with added attention to general issues of interpretation and application of this tool for scientific research.

The IAT is a method for indirectly measuring the strengths of associations among concepts. The task requires sorting of stimulus exemplars from four concepts using just two response options, each of which is assigned to two of the four concepts. The logic of the IAT is that this sorting task should be easier when the two concepts that share a response are strongly associated than when they are weakly associated.

Table 1 presents a schematic describing a typical IAT design for the assessment of association strengths between categories of *men* and *women* and attributes of *good* and *bad*. The IAT consists of seven phases, some of which are practice tasks to acquaint subjects with the stimulus materials and sorting rules. The critical phases of the IAT involve simultaneous sorting of stimulus items representing four concepts (*men*, *women*, *good*, *bad*) with two response options. In one critical phase (B3 and B4 in the example), items representing *men* and *good* (e.g., male faces and words such as wonderful, glorious) receive one response, and items representing the concepts *women* and *bad* (e.g., female faces and words such as terrible, horrible) receive the alternative response. In the second critical phase (B6 and B7 in the example), items representing the concepts *women* and *good* are sorted with one response, and items representing the concepts *men* and *bad* are sorted with the alternative response. For subjects who possess stronger associations of positive evaluation with females compared to males, the second sorting task should be much easier than the first. Likewise, subjects who possess stronger associations of positive evaluation with males compared to females should find the first sorting task to be easier than the second. Ease of sorting can be indexed both by the speed of responding (faster responding indicating stronger associations) and the frequency of errors (fewer errors indicating stronger associations).

In this chapter, we describe the IAT's procedural details, along with evidence for its validity as a measure of association strengths. Also, a variety of threats to validity are identified and correctives are suggested. Finally, we review some of the critical issues concerning the interpretation of IAT effects and some reflections on its potential applications proper use.

Internal Validity

This section reviews issues concerning the internal validity of the IAT including (a) the selection and design of stimulus materials such as category labels and exemplars, (b) procedural features such as the order and length of response blocks, (c) a review of the known extraneous influences on the IAT and potential correctives for those influences if they are available, (d) suggested analytic procedures for the IAT, (e) a review of evidence for the IAT's internal consistency and test-retest reliability, and (f) a review of evidence concerning the fakeability of IAT performance.

Materials

The critical materials of an IAT are four categories defined by category labels (e.g., *men*, *women*, *good*, *bad*) and stimulus items that serve as exemplars for those categories (e.g., male and female faces, and words with good or bad meaning). In most IAT designs, the four categories represent two contrasted pairs, sometimes distinguished as target concepts (e.g., *men-women)* and attribute (e.g., *good-bad*) dimensions.[1] The two dimensions usually define the two nominal features that are of direct interest and create the contrasting identification tasks – e.g., "what is the gender?" for category items, and "what is the evaluation?" for attribute items (Greenwald, Nosek, Banaji, & Klauer, in press). The IAT effect is a comparative measure of the combined association strengths of two associative pairs (*men* with *good*, *women* with *bad*) contrasted with strengths of two other associative pairs (*men* with *bad*, *women* with *good*). In the present case, the resulting score has a relatively simple interpretation as an implicit measure of relative preference for men compared to women.

Design of an IAT requires selecting category labels that define the concepts of interest and stimulus items to represent those concepts. There are some important factors to consider in the selection of these materials.

Ensure that the category membership of stimulus items is clear and used for categorization. The subjects' primary task in the IAT is to identify the category membership of stimulus items as quickly as possible. Each stimulus item must be identifiable as representing just one of the four categories, e.g., *men* or *women* for gender, and *good* or *bad* for evaluation. If the category membership of a stimulus item is difficult to identify or confounded with multiple categories, then subjects may be unable to categorize accurately, or may attempt to complete the task with sorting rules different from those intended for the design.

Task confusion can be reduced by providing multiple cues for identifying the relevant nominal feature of any given stimulus item, so that items clearly represent one and only one of the four categories. For example, confounds between dimensions should be avoided (Steffens & Plewe, 2001). In the current example, using "gendered" *good* and *bad* items such as "nurturing" and "aggressive" could introduce confusion about whether to categorize the items on the basis of gender or evaluation. Also, the distinctiveness of nominal dimensions is enhanced if they use different stimulus modalities, such as faces for gender and words for evaluation, or by using distinct colors or fonts such as gender words in green and evaluation words in white. Finally, strictly alternating response trials between nominal dimensions creates a predictable pattern for the switching between the relevant feature judgments. As an added benefit, alternating trials maximizes task switches, which appear to be important contributors to IAT effects (Klauer & Mierke, 2004).

Another important aspect of exemplar selection is to ensure that stimulus items are categorized on the basis of the intended nominal feature rather than an irrelevant stimulus feature. In other words, it should be difficult to distinguish the two categories of a single nominal dimension (e.g., men or women) using any characteristic except the nominal feature (gender). If the categories *men* and *women* were comprised of Black male and White female faces respectively, category membership would be clear, but subjects could sort items based on race (irrelevant) or gender (relevant). Likewise, if all of the *good* words were more than 10 letters and the *bad* words were less than 5, subjects could sort based on evaluative meaning or length.

<u>Other stimulus characteristics</u>. Stimulus items can be presented as words, pictures, sounds, or in a combination of modalities. Generating stimulus items requires balancing the competing demands of (a) creating an accurate representation of the superordinate category, and (b) avoiding exemplars that are only weakly representative of the category. Nosek, Greenwald, and Banaji (2004) observed that the magnitude and reliability of IAT effects were relatively unaffected by the number of stimulus items per category, except that effects were somewhat weaker when only a single exemplar per category was used. Stimulus sets should contain only items that clearly belong to the target category and, as a group,

represent the intended category in a fashion appropriate for research purposes – for example, not representing a distinct subset of a category (e.g., fruit juices) when the larger category (e.g., soft drinks) is of interest.

Categories in the IAT are constructed as contrasting pairs (*men-women, good-bad*). The resulting IAT score is a relative measure of associations between categories (Greenwald et al., 1998). Whereas the IAT measures relative association strengths involving four categories (*men* with *good* and *women* with *bad* compared to *men* with *bad* and *women* with *good*), one might reasonably be interested in measuring the association of evaluations with men alone (*men* with *good* versus *men* with *bad*). However, as a relative measure, the IAT is not as useful for measuring associations toward single targets (Nosek et al., 2004). The relative measurement feature of the IAT constrains its proper application and interpretation. As a consequence, the selection of comparison categories is of critical importance in design. For research efforts in which single category assessments are of particular interest, a different measure of associations should be considered such as the Go/No-Go Association Task (Nosek & Banaji, 2001) or the Extrinsic Affective Simon Task (De Houwer, 2003).

<u>Procedural Design</u>

Greenwald, Nosek, and Banaji (2003) summarized a standard IAT procedure that requires rapid sorting of exemplars representing two concept categories (*men* and *women*) and two attribute categories (*good* and *bad*) into their nominal categories with a set of seven response blocks (see Table 1): (B1) 20 trials sorting the two target concepts with the same two keys - e.g., 'males' words with the 'e' key, 'females' with the 'i' key, (B2) 20 trials sorting good and bad words using two response keys - e.g., 'good' words with the 'e' key, 'bad' words with the 'i' key, (B3) 20 trials sorting items from all four categories with the same two keys alternating by trial between concept and attribute items - e.g., males and good with 'e', females and bad with 'i', (B4) 40 trials with same sorting rule as B3, (B5) 20 trials of sorting the concept categories with the reverse key mapping from B1 - i.e., 'males' with 'i' key and 'females' with 'e' key, (B6) 20 trials sorting items from all four categories with the opposite key pairings from B3 and B4 - i.e., females and good with 'e' key, males and bad with 'i' key, (B7) 40 trials with same sorting rule as B6. Blocks B3, B4, B6, and B7 comprise the critical data of the task.

In most IAT studies, half of the sample completes the task in the order above, and the other half completes the task with B1, B3, B4 switched with B5, B6, B7. Nosek et al., (2004) proposed changing B5 to 40 response trials as a standard corrective for a persistent extraneous influence of task order (see next section). A comparison of average latency between the first combined sorting condition (B3, B4) and the second (B6, B7) is taken to reveal the relative association strengths between the concepts and attributes. In other words, participants who find it easier to sort men with good (and women with bad) compared to sorting men with bad (and women with good) are said to implicitly prefer males to females.

There are some additional procedural factors that are important and are applied across all response blocks. For each block, the category labels appear on the top left and right of the computer screen to remind participants of the response key mapping rules. When stimulus items are incorrectly categorized, an error indication appears (often a red 'X' immediately below the stimulus item) and the subject is obliged to fix the error by hitting the correct response key before continuing to the next trial. The interval between occurrence of one trial's response and presentation of the following trial's stimulus – the intertrial intervals (ITI) is typically short, although usually not less than 150 ms. Greenwald et al. (1998) reported that use of longer ITIs (up to 750 ms) had no appreciable effect on IAT measures. A relatively short ITI (250 ms may be most often used) allows the measure to be obtained rapidly.

The IAT has been used with procedural variations, usually without any attempt to collect data to discriminate alternate versions. The virtues of the procedure as described here are that it has received very widespread use, achieves satisfactory reliability, and can be administered in just five minutes. Nevertheless, there may be circumstances in which it would be advantageous to alter these procedures. For example, extra practice blocks or trials may be essential for populations that are unfamiliar or unskilled with computers or speeded responding (e.g., the elderly or young children).

Extraneous Influences

Some procedural and person factors that have been shown to have little or no influence on IAT measures include (a) whether a particular category is assigned to the left or right response key (Greenwald et al., 1998), (b) whether the response-stimulus interval (ITI) was as low as 150 milliseconds or as high as 750 milliseconds (Greenwald et al., 1998), (c) whether there is wide variation in the familiarity of stimulus items comprising the attitude object categories (Dasgupta, McGhee, Greenwald, & Banaji, 2000; Dasgupta, Greenwald, & Banaji, 2003; Ottaway, Hayden, & Oakes, 2001; Rudman, Greenwald, Mellot, & McGhee, 1999), and (d) whether subjects were right or left-handed (Greenwald & Nosek, 2001).

On the other hand, there are several extraneous influences that can obscure the measurement of association strengths with the IAT. Procedural and analytic innovations may reduce some of these undesired influences. The following paragraphs summarize the known extraneous influences on the IAT. Empirically identified correctives are offered where available.

<u>Order of combined tasks</u>. The influence of the order of combined tasks mentioned in the previous section is the most commonly observed extraneous factor (Greenwald & Nosek, 2001). Performance of the first combined pairing (B3, B4) tends to interfere with performance of the second (B6, B7). As a consequence, IAT effects is slightly biased toward indicating that the associations drawn upon in the first-performed combined task are stronger than those drawn upon in the second-performed task. Nosek et al. (2004) observed that using 40 response trials in B5 instead of 20 significantly reduced the influence of this extraneous factor. Consequently, increasing the number of trials in B5 is a recommended procedure. In addition, counterbalancing the task order helps to identify and enables statistical removal of the biasing effects of this influence.

<u>Cognitive fluency</u>. Another extraneous influence is individual differences in average response latency, or cognitive fluency. Subjects who perform the task more slowly overall tend to show larger IAT effects than those who perform more quickly (McFarland & Crouch, 2002; Greenwald et al., 2003). This is a common nuisance factor for response latency data. Greenwald and colleagues (2003) introduced a scoring algorithm (D) that, among other improvements, reduces the influence of this factor (see also Cai, Sriram, Greenwald, & McFarland, in press). The scoring algorithm is introduced in the next section.

Researchers may also introduce content-irrelevant response latency tasks to provide a cognitive fluency factor that could possibly be used as the basis for a covariance adjustment.

A particular form of cognitive fluency, task-switching ability, or the facility for switching between judgment tasks (e.g., gender or evaluation), has been extensively examined in the IAT (Mierke & Klauer, 2001, 2003). As with overall speed of responding, the D scoring algorithm appears to significantly reduce the extraneous influence of this factor (Back, Schmukle, & Egloff, in press; Mierke & Klauer, 2003).

Subject age. Perhaps related to effects of variations in cognitive fluency, older subjects tend to show larger IAT effects than younger subjects (Greenwald & Nosek, 2001; Hummert, Garstka, O'Brien, Greenwald, & Mellott, 2002), especially when the original scoring algorithm (Greenwald et al., 1998) is used. The improved scoring algorithm suggested by Greenwald and colleagues (2003) reduces this relationship between age and IAT scores.

Experience with the IAT. Effect magnitudes with the IAT tend to decline with repeated administrations (Greenwald & Nosek, 2001; Greenwald et al., 2003). The D scoring algorithm reduces the influence of this factor, but IAT experience should not be ignored either in designs that use multiple IATs in a single session or that use repeated IAT measures in longitudinal studies with multiple testing sessions. An additional corrective for such designs is to include a control IAT for comparison purposes that is not expected to change as a function of the manipulation or intervention (e.g., Teachman & Woody, 2003).

Order of measures. It is possible that the order in which self-report and IAT measures are completed effect IAT performance as well as self-report. For example, performing self-report measures first may make increase the accessibility of some cognition and affect subsequent IAT performance. Likewise, an IAT that reveals an unexpected association may influence ensuing self-report. The actual effects of task order are not fully understood, though the accumulated evidence suggests that effects in typical circumstances are minimal. In a meta-analysis of IAT studies, Hofmann, Gawronski, Gschwendner, Le, and Schmidt (in press) found no effect of order. And, in experimental manipulations of task order, little to no effect of task order was observed in large web-based samples reported by Nosek et al., 2004. Even so, it is reasonable to suppose that some manipulations of task order will influence IAT effects (see Nosek et al., 2004). A reasonable procedural guideline is to counterbalance order of IAT and self-report measures in the absence of reasons for using just a single order.

Analysis

Greenwald and colleagues (2003) evaluated a variety of candidate scoring algorithms on a wide range of psychometric criteria (sensitivity to known influences, correlations with parallel self-report measures, internal consistency, and resistance to extraneous procedural influences) on very large Internet samples. The best performing algorithm (D) strongly outperformed the conventional scoring procedures and was recommended by Greenwald and colleagues (2003). The D algorithm has since been shown to have additional psychometric benefits over the conventional scoring procedures (Back, et al., in press; Cai, et al., in press; Mierke & Klauer, 2003).

The algorithm recommended by Greenwald et al., (2003) has the following steps for IAT designs in which subjects must correct errant responses before continuing: (1) use data from Blocks 3, 4, 6, and 7 (see Table 1); (2) eliminate trials with latencies > 10,000 ms; (3) eliminate subjects for whom more than 10% of trials have latencies <300 ms; (4) compute one standard deviation for all trials in Blocks 3 and 6, and another standard deviation for all trials in Blocks 4 and 7; (5) compute means for trials in each of the four blocks (3, 4, 6, 7); (6) compute two difference scores (one between 3 and 6 and the other between 4 and 7) subtracting what is intended to represent the high (positive) end of the measure from the block containing associations representing the low end; (7) divide each difference score by its associated standard deviation from Step 4; and (8) average the two quotients from Step 7.[2]

Reliability

A persistent challenge for implicit measures is to achieve substantial internal consistency and test-retest reliability. For example, some evaluative priming measures show weak internal consistency (e.g., split-half $r = .06$, Olson & Fazio, 2003; see also Bosson, Swann, and Pennebaker, 2000; Fazio & Olson, 2003). Early variants of the Go/No-go Association Task (GNAT) show relatively weak reliability when its signal detection method variation is used (average split-half $r = .20$; Nosek & Banaji, 2001). In a direct comparison, the IAT outperformed the Extrinsic Affective Simon Task (EAST $\alpha = .19, .24, .19$; IAT $\alpha > .75$; Teige, Schnabel, Banse, & Asendorpf, 2004). Part of the IAT's acceptance as an implicit measure may be attributable to its achieving greater reliability than other latency-based implicit measures.

Internal consistency. The IAT has displayed satisfactory internal consistency, which is relatively rare for latency-based measures. For example, Bosson et al. (2000) observed an split-half internal consistency for the self-esteem IAT of $r = .69$ compared to *r*'s of -.05 to .28 for other latency-based implicit self-esteem measures. Internal consistency estimates (split-half correlations or alphas) for the IAT measures tend to range from .7 to .9 (Greenwald & Nosek, 2001; Schmukle & Egloff, 2004).

Test-retest reliability. Another form of measurement consistency is test-retest reliability – the consistency of measurement over time. High test-retest correlations should occur to the extent that the IAT is a trait measure rather than a state measure. If the IAT is state-dependent, then test-retest reliability may be low even when internal consistency estimates are high. Egloff, Schmukle, and colleagues have conducted the most thorough tests of internal consistency and test-retest reliability of IAT measures (Egloff, Schwerdtfeger, & Schmukle, 2005; Schmukle & Egloff, in press, 2004). Figure 1 is based on a summary table of IAT test-retest studies contained in Schmukle and Egloff (2004) and including a few additional studies. The x-axis presents the time, in days, between two administrations of the IAT, and the y-axis is the correlation between the two tests. Across studies, the IAT shows stable test-retest reliability (median $r = .56$) that varies little with retest interval. Two qualifications of this conclusion are (1) only one study has examined test-retest reliability with more than a 1 month gap (1 year; Egloff, et al., 2005); and (2) Figure 1 combines data from a variety of tasks (anxiety, racial attitudes, extroversion) – possible variation in test-retest reliability by content domain is undetermined (Schmukle & Egloff, in press). Even so, the effect of time between tests on test-retest reliability is unaffected by the presence of the outlier study. Schmukle and Egloff (2004) concluded that the IAT has satisfactory test-retest reliability while also showing evidence of both trait-specific, an individual difference that is stable across time, and occasion-specific variation.

Fakeability

All psychological measures attempt to assess some aspect of mental content. This typically requires the willingness of the respondent to be assessed, and the honest efforts of the respondent to answer accurately or otherwise follow instructions. Direct measures make the meaning of the response plain, and allow the respondent to straightforwardly determine the response content. Indirect methods attempt to reduce the likelihood of deliberate faking by obscuring what is being measured, how it is being measured, or limiting the ability to control the response content. In this regard, implicit measures comprise a subset of indirect methods.

Investigations of IAT fakeability across multiple content domains including shyness, extraversion, moral identity, attitudes toward flowers versus insects, attitudes toward sexual orientation, and attitudes toward racial groups collectively suggest that (a) the IAT is much less fakeable than self-report, (b) the IAT is not very fakeable when subjects are given only abstract instructions to do so (e.g., "try not to appear shy") and, (c) two factors, experience with the IAT and explicit instructions about how to control IAT scores, increase fakeability (Asendorpf, Banse, & Mucke, 2002; Asendorpf, Banse, & Schnabel, 2003; Banse, Seise, & Zerbes, 2001; Egloff & Schmukle, 2002; Kim, 2003; Perugini & Leone, 2004; Steffens, in press). Also, the IAT often reveals associations that subjects do not endorse, or would prefer not to reveal suggesting that it is resistant to deliberative alteration in practice. For example, most White subjects show a consistent implicit preference for Black relative to White despite an explicit desire not to do so, and many Black subjects do not show an implicit preference for Black relative to White despite an explicit desire to do so (e.g., Nosek, Banaji, & Greenwald, 2002).

The fact that IAT measures are often only weakly correlated with self-reported attitudes suggests that deliberate faking may not be substantial under typical study conditions. Even so, this role of faking in IAT performance deserves further attention. Another fertile, but untapped question in this area is whether faking, when it does occur, can be empirically distinguished from honest task performance. It is possible that algorithms could be designed to distinguish actual from faked IAT data.

The issue of faking is related to the possibility that subjects could attempt to exert cognitive control in order to either suppress or overcome their automatic associations. This type of control may not be faking in the sense of trying to create a false impression (i.e., people may genuinely reject their automatically activated thoughts), but it suggests that cognitive control may mask automatic associations that are otherwise activated. In other words, response alternation in the IAT might occur by deliberate alteration of the task procedures, but may also occur through deliberate effort to alter one's mind. For example, Akalis and Banaji (2004) have shown that instructions to 'think good thoughts' or 'think compassionately' reduce bias toward overweight persons and in a novel group context. Conrey and colleagues have developed a promising multinomial model for parceling the various controlled and automatic cognitive processes that contribute to IAT effects (Conrey, Sherman, Gawronski, Hugenberg, & Groom, in press). This modeling approach has the virtue of highlighting the fact that the IAT, like all tasks, is not process pure, and that the component processes involved in IAT effects such as automatic influences and efforts to overcome bias may be distinguishable through sophisticated experimental and statistical methods.

Construct Validity

This section reviews evidence for the construct validity of the IAT focusing on (a) the relationship between the IAT and other implicit measures, (b) the relationship between the IAT and parallel self-report measures, (c) the predictive validity of the IAT, (d) evidence for independent variable influences on the IAT, and (e) evidence concerning the development of IAT effects in children.

The relationship between IAT and other implicit measures

The IAT is one of a diverse family of measures that are referred to as *implicit*. In one of the few investigations comparing multiple measures, Bosson, Swann, and Pennebaker (2000) observed weak relations among seven implicit measures of self-esteem, including the IAT (*r*s ranged from -.14 to .23). Also, a number of studies have compared the IAT with variations of evaluative priming and found weak relations (Bosson et al., 2000; Marsh, Johnson, & Scott-Sheldon, 2001; Olson & Fazio, 2003; Sherman, Presson, Chassin, Rose, & Koch, 2003).

There are two factors that appear to contribute to the observation of weak relations among implicit measures. First, implicit measures often demonstrate relatively weak reliability compared to other forms of psychological measurement. Reliability of measures set upper limits on their possible correlations with other measures. For example, the maximum, meaningful correlation that can be observed between a measure with reliability of .10 and a measure with perfect reliability (1.0) is .32, which is estimated by calculating the product of the square roots of the two reliability coefficients (Nunnally & Bernstein, 1994, p. 241). If the second measure's reliability is .50 instead of perfect, the maximum observable correlation is only .22. It is not uncommon for response-latency based measures to show reliabilities well below .50. For example, Bosson and colleagues (2000) reported internal consistencies (α's) for the IAT, supraliminal priming, subliminal priming, and stroop of .88, -.16, .49, and -.38 respectively. Further, of the response latency methods, only the IAT showed test-retest reliability greater than $\underline{r} = .25$ (IAT test-retest $\underline{r} = .69$; Bosson et al., 2000). In sum, the relations among implicit measures (and between implicit measures and other variables) will be underestimated to the extent that they are unreliably assessed.

When unreliability is accounted for in models, stronger relations emerge. Cunningham, Preacher, and Banaji (2001) used structural equation modeling to estimate disattenuated correlations between implicit measures and observed correlations ranging from .53 to .77 between two versions of a racial attitude IAT and a response window evaluative priming measure. Also, Nosek and Banaji (2001) observed a disattenuated correlation of .55 between the IAT and the GNAT.

While some of this weak relationship among implicit measures is surely attributable to low reliability, the relations may also reflect heterogeneity of cognitive processes that contribute to the various measures. The term *implicit* has become widely applied to measurement methods for which subjects may be unaware of what is being measured, unaware of how it is being measured, or unable to control their performance on the measure. Identification of the cognitive processes that contribute to different measures will promote a more nuanced description and categorization of methods based on the particular processes that they engage.

Some research efforts have identified relevant differences in measurement methods to clarify the relations among implicit measures (e.g., Brauer, Wasel, & Niedenthal, 2000; Olson & Fazio, 2003). Olson and Fazio (2003), for example, commented on the fact that evaluative priming allows spontaneous categorization of target concepts (primes), while the IAT constrains concept categorization to a particular feature of the stimulus items (e.g., the race of face). In support of this spontaneous versus constrained process distinction, a variation of the priming measure that encouraged subjects to categorize primes in racial terms showed stronger correspondence with the IAT than did the more typical priming procedure.[3]

The Relationship between IAT and Self-Report

Some of the first research efforts with the IAT emphasized the distinctiveness of implicit and explicit cognition in finding weak to absent relations between implicit measures, like the IAT, and self-report (Greenwald et al., 1998). More recent research has shown that, in some cases, the IAT and self-report can be strongly related (Greenwald et al., 2003; Hofmann et al., in press; Nosek et al., 2002; Nosek, 2005). At the most extreme, a large Internet sample of data measuring preferences for Al Gore relative to George Bush revealed a disattenuated correlation of .86 with an explicit measure of candidate preference (Greenwald et al., 2003). In a meta-analysis of IAT and self-report correlations, Hofmann and colleagues (in press) reported an average *r* of .24, and in an investigation of 57 different content domains, Nosek (2005) reported an average correlation of .37 (when similar data were subjected to structural equation modeling, the disattenuated *r* was .46; Nosek & Smyth, 2004).[4]

Convergent and Discriminant Validity

The realization that the IAT and self-report are related introduces important questions about whether they measure distinct constructs. In a multitrait-multimethod (MTMM) investigation of the IAT and self-report across seven attitude domains, Nosek and Smyth (2004) found strong evidence for both convergent and discriminant validity – IAT attitude measures were related to their corresponding self-report measure and not measures of other traits. Further, using structural equation modeling, this MTMM investigation revealed that the best-fitting models represented the IAT and self-report as related, but distinct constructs, rather than as a single attitude construct, even after accounting for common method variance in both measures (Nosek & Smyth, 2004; see also Cunningham, Nezlek, & Banaji, 2004). This extends similar findings for individual constructs such as Greenwald and Farnham (2000) for self-esteem, and Cunningham et al. (2001) for racial attitudes. Finally, Nosek (2005) reported evidence that the relationship between the IAT and self-report is moderated by multiple interpersonal (self-presentation, perceived distinctiveness from the norm) and intrapersonal (e.g., evaluative strength) features of attitudes.

Predictive Validity

Evidence for the predictive validity of the IAT is emerging from a wide-variety of domains. As already reviewed, IAT scores are predictive of self-report attitudes and the strength of that relationship is moderated by multiple factors (Hofmann et al., in press; Nosek, 2005). Poehlman, Uhlmann, Greenwald, and Banaji (2004) recently compiled 61 studies with 86 individual effect sizes in which the predictive validity of the IAT when perceptual, judgment, and action processes were examined as criterion variables. From the meta-analysis the authors draw two main conclusions. In studies that involve some measure of discrimination toward a social group, both explicit and IAT measures predict behavior but the IAT does a superior job of prediction (mean $r_{IAT} = .25$, mean $r_{self\text{-}report} = .13$). In studies that measure brand preferences or political candidate preferences, both IAT and explicit measures predict the outcome, but explicit measures do a superior job of prediction (mean $r_{IAT} = .40$, mean $r_{self\text{-}report} = .71$).

Malleability and Development

Conceptions of automaticity have often emphasized its consistency and inflexibility. This feature of automaticity has undergone a slow revision with the realization that automaticity is conditional and malleable on features of the present context (Gilbert & Hixon, 1991; Kahneman & Treisman, 1984; Macrae, Bodenhausen, Milne, Thorn, & Castelli, 1997; Wittenbrink, Judd, & Park, 2001). Evidence for conditional automaticity in implicit cognition has blossomed in recent years in investigations of the malleability of attitudes, identity, and beliefs (Blair, 2002; Blair, Ma, & Lenton, 2001; Dasgupta & Asgari, 2004; Dasgupta & Greenwald, 2001; Ferguson & Bargh, 2003; Florack, Bless, & Piontkowski, 2003; Foroni & Mayr, in press; Lowery, Hardin, & Sinclair, 2001; Mitchell, Nosek, & Banaji, 2003; Richeson & Ambady, 2003; Richeson & Nussbaum, 2004; Rudman, Ashmore, & Gary, 2001; Teachman

& Woody, 2003; Wittenbrink et al., 2001). For example, Lowery and colleagues (2001) demonstrated that implicit racial bias, measured by the IAT and subliminal priming, was substantially weaker when the experiment was administered by an African American compared to a Caucasian American. Likewise, Foroni and Mayr (in press) showed that attitudes toward flowers relative to insects differed as a function of reading a short story about 'dangerous' or 'good' flowers.

Development of implicit cognition. Baron and Banaji (2004) devised a child friendly version of the IAT, called the Ch-IAT (available at www.people.fas.harvard.edu/~banaji). Using this version, children as young as age 4 have been tested on race and gender attitudes. Grouping white participants into three categories of age 6, age 10, and adults, Baron and Banaji (2004) found that the IAT race attitude remains constant across the three age groups. Explicit attitudes toward Black Americans change systematically with 10 year olds reporting lower race bias than 6 year olds and adults showing no race bias at all. Dunham, Baron, and Banaji (2004) have also studied Japanese children and adults in small town of approximately 6000 residents, and found that children prefer their own group (Japanense) compared to both Whites and Blacks, whereas adult Japanese show less liking of Japanese when contrasted with Whites than when contrasted with Blacks. The origins of bias, and implicit cognition more generally, is likely to become a more active area of research for those interested in the early stages of social category knowledge, preference formation, and the extent to which such categories are or are not "essential" and part of core knowledge begin to incorporate implicit measurement into their research.

Malleability and Fakeability. There is an intriguing contrast between findings that relatively simple situational manipulations can result in shifts in automatic evaluations, and findings that spontaneous faking of IAT effects is not so easy. Consider, for example, Kim's (2003) observation that instructing subjects to fake the race attitude IAT resulted in no difference in effects compared to a control group, whereas Lowery et al. (2001) showed a dramatic shift in automatic racial evaluations just by varying the race of the experimenter. Even more dramatically, Foroni and Mayr (in press) had participants complete a flower-insect attitude IAT twice, once after reading a 'pro-flowers' story and once after receiving one of three task manipulations (a) read a brief fantasy story about dangerous flowers and valuable insects following a nuclear war, (b) instructions to associate insects with negative and flowers with positive, or (c) instructions that the IAT is a lie-detector and that they should try to deceive it by pretending that they dislike flowers and like insects. Only the first of these three conditions elicited less positivity toward flowers relative to insects compared to the control condition. Remarkably, this suggests that it is easier to shift IAT effects through indirect means like telling a story than through a request to deliberately alter the effects. This demonstration underscores the malleability of automatic cognition, and suggests limitations of the involvement of deliberative processes in producing that malleability.

Interpretation of IAT effects

This section considers some of the interpretive issues that regularly arise in relation to IAT effects, including (a) the IAT as a relative measure of association strengths, (b) the influence of stimulus items versus category labels on IAT effects, (c) whether IAT effects should be considered more accurate or 'real' than self-report, and (d) the cognitive processes and neurological correlates of IAT effects.

Relative Measurement

A prior section pointed out that the IAT is a relative measure of association strengths (Greenwald et al., 1998). Some researchers have attempted to circumvent this procedural constraint of the IAT by applying analytic methods designed to measure absolute associations that analyze response latencies for only those trials in which an exemplar from just one of the two target concepts was presented (e.g., Baccus, Baldwin, & Packer, 2004; de Jong et al., 2001; Gemar, Segal, Sagrati, & Kennedy, 2001). For example, in an attitude IAT contrasting Black faces from White faces, response latencies to categorizing Black faces might be extracted from the two conditions in an effort to measure liking for Blacks irrespective of evaluations of Whites. However, the individual response trials in the IAT are not independent events, and these analytic methods do not isolate single associations from the IAT (Nosek et al., 2004).

Another approach for measuring absolute associations with the IAT is to contrast the target category (e.g., self) with a neutral category (e.g., furniture, middle, animals, shapes). This type of application assumes that the neutral contrasting category contributes no meaningful variability to measurement and thus results in a score that can be interpreted as an absolute assessment of the target concept (e.g., Jajodia & Earlywine, 2003; Sherman et al., 2003). This approach is viable to the extent that the contrasting category is truly neutral and produces no meaningful variability in measurement. While this type of contrast category selection may reduce meaningful variability contributed by the contrast, it is not ideal for measuring single associations because of the strong assumptions that are required for interpretation.

Another strategy that has been used to measure single associations with the IAT is a 'one category' variation (Wigboldus, in press). In this version of the IAT, only three categories are used instead of four (one target concept and two attribute concepts). The two critical conditions of the task involve categorizing two sets of concepts with one response and one concept with the alternate response (e.g., *Black*+*bad* to the left, *good* to the right; then *bad* to the left and *Black*+*good* to the right). The simplicity of this approach is appealing. However, our own parameter testing with this IAT variant suggested a substantial threat to internal validity (Nosek & Banaji, 1997). Some subjects appeared to spontaneously or deliberately simplify the task to search for items belonging to the unpaired attribute category and accept or reject all items based on this concept exclusively (i.e., if *bad* hit the key, otherwise hit the other key), thus reducing attention to the nominal features of the (single) target concept items included in the task. Perhaps additional procedural innovations to the one-category task can reduce these threats to internal validity, but those innovations have not yet emerged.

A final alternative for measuring single associations is to avoid the IAT altogether and instead use a measure that is designed to assess them. Two measures that may provide this flexibility include the Go-No go Association Task (Nosek & Banaji, 2001) and the Extrinsic Affective Simon Task (De Houwer, 2003). The psychometric properties of these new measures are not as well understood as the IAT, so their usefulness as measurement methods remains to be determined.

Attitudes toward Stimulus Items or Category Labels

A feature that clearly differentiates the IAT from its implicit cousin, evaluative priming (Fazio, Sanbonmatsu, Powell, & Kardes, 1986; Wittenbrink et al., 1997), is that the IAT requires explicit categorization of target stimuli into superordinate concepts. In sequential priming measures, primed concepts are not necessarily explicitly categorized and, in some cases, may not even be consciously perceived (Draine & Greenwald, 1998). The IAT requires that the subject explicitly categorize stimulus items into a specified superordinate category. This difference may have important effects on measurement such that participants may idiosyncratically categorize items in sequential priming procedures, but must arrive at a specific categorization for IAT performance (Olson & Fazio, 2003).

This distinctive procedural feature has produced interest in the extent to which IAT effects are influenced by the superordinate categories defined by the category labels, and the individual features of the stimulus exemplars. Some researchers have assumed that the IAT effects are purely a result of stimulus features (Brendl, Markman, & Messner, 2001; Mitchell, 2004), and others have concluded that IAT effects are determined largely by the category labels (De Houwer, 2001; Fazio & Olson, 2003).

The available evidence suggests that the answer is in between. Category labels appear to be critical for constraining the interpretation of the stimulus items. At the same time, the stimulus items, as a set, can affect the construal of the target category (Govan & Williams, 2004; Mitchell, et al., 2003; Nosek et al., 2004; Steffens & Plewe, 2001). For example, males' automatic association between romantic fantasies and their partner varied as a function of whether the romantic fantasies were sexualized stimulus items (e.g., vixen) or not (e.g., Cinderella; Rudman & Heppen, 2003, Studies 2 and 3). Also, in a Gay-Straight attitude IAT, changing just two stimulus items in the 'Gay people' category from representations of gay women to gay men resulted in stronger pro-straight preferences in the latter compared to the former representation (Nosek et al., 2004). Finally, the importance of the category labels for constraining interpretation of stimulus items is made plain by unpublished parameter testing from our laboratories in which math and arts stimulus items were replaced with meaningless symbols (X's and O's; Nosek & Banaji, 1997). Instructing subjects on the proper interpretation of the symbols (X means math, O means arts) was sufficient to elicit IAT effects consistent with 'real stimuli' IATs.

In sum, IAT design requires careful attention to the selection of both category label and stimulus items. Category labels are clearly of great importance the IAT, but the stimulus exemplars can nevertheless influence the construal of those categories. Stimulus exemplars can aid in the definition of the superordinate category (e.g., whether gay people refers to gay men, lesbians, or both).

Does the IAT Reveal Cognitions that are More 'True' or 'Real' than Self-Report?

A rarely asserted interpretation of the IAT is that it might serve as a lie-detector, revealing associations that are more 'real', 'true', or accurate than self-report. Our review of the IAT literature has not found any article that endorsed this position, but we did find a number of articles that criticized users of the IAT for espousing that position, either incorrectly attributing the lie-detector view to the originators of the IAT, or attributing the view without supporting citation (e.g., Arkes & Tetlock, in press; Gehring, Karpinski, & Hilton, 2003; Karpinski & Hilton, 2001).

To the extent that the IAT assesses implicit cognition as defined by Greenwald and Banaji (1995) and discussed in the opening section of this paper, it reveals associations that an individual may not want to report, and may not be aware of possessing. So, IAT and self-report can differ because (a) the individual is unaware of the implicitly measured associations, and uses introspection to generate a unique explicit response, (b) the individual is aware of the implicitly measured associations, but genuinely rejects them as not conforming to his or her belief system and so reports a distinct explicit response, or (c) the individual is aware of the implicit associations, but chooses to report an alternative explicit response due to social concern about the acceptability of such a response. Only the third case would fit with the conception of detecting a deliberate hiding of privately endorsed evaluations. In all three cases, implicit and explicit assessments may have separate predictive utility, which would indicate that both are 'real' measures.

If the IAT measured only associations that fit into category (c) above, then the notion of a lie detector might have merit. However, the evidence is clearly against such an understanding of the IAT. Some of those who complete an IAT that reveals something undesirable (e.g., racial bias) are surprised by the result and may have feelings of guilt (Monteith, Ashburn-Nardo, Voils, & Czopp, 2002; Monteith, Voils, & Ashburn-Nardo, 2001). The IAT clearly should not be regarded simply as a measure of the constructs assessed by self-report, but with all self-presentation concern removed. Self-presentation (for genuine or deceptive reasons) is just one of a variety of factors that moderate the relationship between implicit and explicit measures (Nosek, 2005).

A second way in which IAT results might be considered more 'real' than self-report is in the ability of each to predict psychological outcomes. If the IAT is consistently superior in outcome prediction, then it might be considered more 'real' with regard to its superior predictive validity. However, in a meta-analysis of predictive validity Poehlman et al. (2004) reported that the IAT and self-report each have domains of superior predictive validity.

In sum, it does not appear useful to classify the IAT or self-report as having distinguishable degrees of access to *reality* or *truth*. Each is a real assessment – one intended to measure products of introspection, the other not dependent on introspection.

What Processes are Involved in IAT Effects?

Process models. The IAT has enjoyed a period of sustained empirical use during which its creators, developers, and users have remained relatively calm about the absence of an established cognitive model of performance at the task that generates the IAT measure. There have been a few notable attempts to interpret at least portions of the cognitive processes involved in performing the IAT (Brendl, Markman, & Messner, 2001; Conrey et al., in press; De Houwer, 2001; Greenwald, Nosek, Banaji, & Klauer, 2005; Hall, Mitchell, Graham, & Lavis, 2003; Mierke & Klauer, 2001, 2003; Olson & Fazio, 2003; Rothermund & Wentura, 2001, 2004). The points that seem well established are (a) that the IAT involves representations at the level of categories more strongly than those at the level of the category's exemplars (esp. De Houwer, 2001) and (b) that the difficulty of switching between the IAT's two discrimination tasks when two concepts assigned to the same key are weakly associated contributes substantially to a slowing of performance that plays an important role in the measure (esp. Mierke & Klauer, 2001).

Having a more comprehensive model of the IAT's performance would likely benefit research using the IAT, possibly by affording greater efficiencies in administration, by suggesting design changes that would increase the IAT's construct validity as a measure of association strength, or by making further progress toward developing effective non-relative measures of association strength. Because there are now several laboratories working on various aspects of deciphering the cognitive mechanisms of the IAT, it may be reasonable to expect important progress in this area before the IAT reaches age 10.

Neurological correlates of IAT effects. The first social cognitive neuroscience study involving the IAT (Chee, Sriram, Soon, & Lee, 2000) identified brain regions engaged when subjects were immersed in performing the IAT itself. This is an interesting way to motivate the question of the mechanism underlying the IAT because we already have information on the processes that particular brain regions are involved in. The most satisfactory evidence from this study is the finding that the regions that are active during the IAT procedure are similar if not isomorphic with those involved in the Stroop task: the anterior cingulate, ventrolateral PFC, and dorsolateral PFC.

Phelps et al. (2000), compared the IAT to amygdala activation, a brain region associated with fear or negative emotional responding, by having participants view black and white faces while being scanned, making a L or R judgment on each face. Relative Black > White activation was found to be correlated with an IAT measure of race bias and a black > white startle eyeblink response. No such correlations were observed using the Modern Racism Scale.

Cunningham, et al. (in press) showed that the correlation between the imaging data and the IAT are much stronger when the faces are presented subliminally than supraliminally. These data also gave signs of a dampening down of the subcortical response to black faces in those who showed greater dorsolateral PFC, ventrolateral PFC, and anterior cingulated cortex, regions known to be involved in inhibition, conflict resolution, and control. Such evidence provides converging support for the idea that IAT performance is connected to the response of an early detection emotion module whose activity is also likely to be consciously controlled via activity in regions that kick in later based on more deliberative thought.

Are the associations measured by the IAT available to introspection? The literature on the IAT appears to reach consensus that the IAT bypasses introspective access, and that IAT effects are influenced by automatic processes (though the degree to which the IAT is influenced by automatic processes is still uncertain). To what extent does the IAT capture those associations that are not available to introspection (Fazio & Olson, 2003)? That is, in what sense (if any) can IAT effects be said to reflect cognitive processes that exist outside of awareness?

It is certainly the case that subjects can be aware of their performance on the IAT after completing it (Ashburn-Nardo et al., 2001) and possibly as they are completing it. Also, it is clear from subject reports that some (though not all) are aware of what the task is intended to measure while they are performing it. However, neither of these senses of awareness address whether introspection is involved in producing the end result that is the IAT score.

The question of self-awareness is difficult to answer definitively because it is not possible to assess the contents of one's awareness except through self-report. Any null result could mean that the questions were not posed appropriately. Even so, the evidence suggests that the IAT can measure associations that can escape some layers of awareness. For many content domains, subjects (including the authors of this paper) were genuinely surprised by the outcome of their IAT performance. In many cases, the task performance seems so distant from expectation (presumably based on introspection) that subjects rapidly generate alternative explanations for the result that are external to the person and are often known to be incorrect, such as appeal to known-incorrect hypotheses about the order in which combined tasks were done, the sides on which stimuli were located, handedness, individual differences in eye-hand coordination, and the familiarity of stimulus items.

In controlled investigations, Mitchell et al. (2003) gave subjects experience with the IAT and then asked them to predict their performance for a novel domain. Predictions were unrelated to actual performance. Finally, in unpublished studies from our laboratories, subjects with and without IAT experience, and with and without background information about the IAT, have consistently failed to predict task performance.

Justifiable Applications of the IAT

The accumulated evidence for the construct and predictive validity of the IAT in assessing individual differences shows that it is thriving as a research tool and will likely continue to do so. The

IAT's successes have also prompted interest in applying it in diagnostic and selection settings. Such potential applications should be approached cautiously with careful attention to acceptable interpretations of IAT effects. The known malleability of the IAT implies that its predictive validity is moderated by situational variables. Further, the known potential for controlled processes to override automatic processes suggests that increased controlled processing can impair the IAT's predictive correlations. Finally, the fact that the IAT and self-report both have spheres of superior predictive validity indicates that the IAT is not properly interpreted as a lie detector or as revealing something more *true* or more *real* than self-report.

Until understanding of the IAT's predictive validity develops further, it is premature to use the IAT as a diagnostic indicator for conclusions that have important, direct, and personal consequences - for example, as a device for selection for employment. Applications that reach beyond what can be justified by available evidence may backfire by producing public or professional reactions that can retard the orderly progress of discovery.

The IAT's best current applications are in education, where it has been used to afford insight into automatic associative processes that are introspectively inaccessible. The IAT is also beginning to be used as a clinical diagnostic tool. The IAT's psychometric properties are currently adequate for this type of clinical use in research settings. However, for more sensitive diagnostic tasks it is more appropriate - pending further research development - to treat the IAT as a useful adjunct to diagnosis than to treat it as a self-sufficient procedure. This is not to disparage the compelling evidence for the validity of the IAT, but to point out that there is still much to learn before its appropriate applications are known.

Conclusion

In its seventh year, the IAT is showing a rapid growth in maturity with a solid base of evidence for its internal, construct, and predictive validity. Still, there are many issues unresolved about the nature of the IAT, and its potential for revealing disquieting aspects about human minds. With the vigorous on-going research programs testing the limits of the IAT from many directions, the IAT is a tool worthy of continued scrutiny. As research in implicit cognition continues to grow and age, insights with the IAT may even lead to progeny that will reach beyond the IAT's own capabilities.

Footnotes

[1] The concept-attribute designation is not always true for the IAT. More generally, the IAT involves measurement of associations among four concepts.

[2] A nearly equivalent alternative for IAT designs in which errant trials are not corrected by the subject has the following steps: (1) use data from Blocks 3, 4, 6, and 7 (see Table 1); (2) eliminate trials with latencies > 10,000 ms; (3) eliminate subjects for whom more than 10% of trials have latencies <300 ms; (4) computer the mean of correct latencies for each of the four blocks; (5) replace each error latency with the block mean from step 4 + a 600 millisecond error penalty; (6) compute one standard deviation for all trials in Blocks 3 and 6, and another standard deviation for all trials in Blocks 4 and 7; (7) compute means for trials in each of the four blocks (3, 4, 6, 7); (8) compute the two difference scores; (9) divide each difference score by its associated standard deviation from Step 6; and (10) average the two quotients from Step 9. Table 4 in Greenwald et al. (2003) was confusing with steps 5 and 6 above being presented in reversed order for this form of the algorithm. The consequences of switching the order of 5 and 6 (SD calculated with or without error corrected trials) are very minor, but the present algorithm has the advantage of creating specific boundary conditions [-2, +2] of possible IAT D scores (Sriram & Greenwald, 2005).

[3] While plausible, the definitiveness of this finding is tempered by the fact that the typical and modified versions of the evaluative priming procedure also had different reliabilities (typical priming split-half r = .04; 'category' priming split-half r = .39). Taking those values, which probably underestimate the reliability of the whole measure, the maximum possible correlations with the IAT (split-half r = .53) in that study were just r = .15 for the typical priming measure and r = .45 for the category priming measure.

[4] The difference between the Hofmann et al. (2004) meta-analysis and the Nosek (2005) finding can be understood by noting the difference in domains examined. Hofmann et al. examined the existing literature, which has a heavy emphasis on content domains that are likely to elicit weak correlations (e.g., racial attitudes) whereas Nosek sampled attitudes across a wide variety of domains, many of which showed moderate to substantial correlations.

References

Akalis, S., & Banaji, M. R. (2004). Modification of the IAT effect via compassion. Unpublished data, Harvard University.

Arkes, H., & Tetlock, P. E. (in press). Attributions of implicit prejudice, or "Would Jesse Jackson 'fail' the Implicit Association Test?" Psychological Inquiry.

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

Asendorpf, J. B., Banse, R., & Schnabel, K. (2003). Fakability of an Implicit Association Test (IAT) and a new Implicit Association Procedure (IAP) for shyness. Unpublished manuscript.

Ashburn-Nardo, L., Voils, C., & Monteith, M. (2001). Implicit associations as the seeds of intergroup bias: How easily do they take root? Journal of Personality and Social Psychology, 81(5), 789-799.

Back, M. D., Schmukle, S. C., & Egloff, B. (in press). Measuring task-switching ability in the Implicit Association Test. Experimental Psychology.

Baccus, J. R., Baldwin, M. W., & Packer, D. J. (2004). Increasing implicit self-esteem through classical conditioning. Psychological Science, 15(7), 498-502.

Banse, R., Seise, J., & Zerbes, N. (2001). Implicit attitudes toward homosexuality: Reliability, validity, and controllability of the IAT. Zeitschrift fur Experimentelle Psychologie, 48, 145-160.

Bargh, J. A. (1997). The automaticity of everyday life. In Wyer, R. S. Jr. (Ed). Advances in Social Cognition Vol.10, 1-61.

Baron, A. S., & Banaji, M. R. (2004). The development of implicit attitudes: Evidence of race evaluations from ages 6, 10, and adulthood. Manuscript submitted for publication.

Blair, I. V. (2002). The malleability of automatic stereotypes and prejudice. Personality and Social Psychology Review, 6, 242-261.

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

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

Brauer, M., Wasel, W., & Niedenthal, P. (2000). Implicit and explicit components of prejudice. Review of General Psychology.

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

Cai, H., Sriram, N., Greenwald, A. G., & McFarland, S. G. (in press). The Implicit Association Test's D measure can minimize a cognitive skill confound: Comment on McFarland and Crouch (2002). Social Cognition.

Chee, M.W.L., Sriram, N., Soon, C. S., & Lee, K. M. (2000) Dorsolateral prefrontal cortex and the implicit association of concepts and attributes. Neuroreport: For Rapid Communication of Neuroscience Research, 11(1), 135-140.

Conrey, F. R., Sherman, J. W., Gawronski, B. Hugenberg, K., & Groom, C. (in press). Beyond automaticity and control: The quad-model of behavioral response. Journal of Personality and Social Psychology.

Cunningham, W.A., Johnson, M.K., Raye, C.L., Gatenby, J.C., Gore, J.C., & Banaji, M.R. (in press). Dissociated Conscious and Unconscious Evaluations of Social Groups: An fMRI Investigation. Psychological Science.

Cunningham, W. A., Preacher, K. J., & Banaji, M. R. (2001). Implicit attitude measures: Consistency, stability, and convergent validity. Psychological Science, 12(2), 163-170.

Cunningham, W.A., Nezlek, J.B., & Banaji, M. R. (2004). Implicit and explicit ethnocentrism: Revisiting the ideologies of prejudice. Personality and Social Psychology Bulletin, 30(10), 1332-1346.

Dasgupta, N. & Asgari, S. (2004). Seeing is believing: Exposure to counterstereotypic women leaders and its effect on the malleability of automatic gender stereotyping. Unpublished manuscript.

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

Dasgupta, N., Greenwald, A.G., Banaji, M.R. (2003). The first ontological challenge to the IAT: Attitude or mere familiarity? Psychological Inquiry, 14(3&4), 238-243.

Dasgupta, N., McGhee, D. E, Greenwald, A. G., & Banaji, M. R. (2000). Automatic preference for White Americans: Eliminating the familiarity explanation. Journal of Experimental and Social Psychology, 36, 316-228.

De Houwer, J. (2001). A structural and process analysis of the Implicit Association Test. Journal of Experimental Social Psychology, 37, 443-451.

De Houwer, J. (2003). The extrinsic affective Simon task. Experimental Psychology, 50(2), 77-85.

de Jong, P. J, Pasman, W, Kindt, M, & van den Hout, M. A. (2001). A reaction time paradigm to assess (implicit) complaint-specific dysfunctional beliefs. Behaviour Research & Therapy, 39(1), 101-113.

Draine, S. C., & Greenwald, A. G. (1998). Replicable unconscious semantic priming. Journal of Experimental Psychology: General, 127(3), 286-303.

Dunham, Y., Baron, A., & Banaji, M. (2004). Developmental social psychology: Outlining a new approach to the study of prejudice in children. Jean Piaget Society, Toronto, Canada

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

Egloff, B., Schwerdtfeger, A., & Schmukle, S. C. (2005). Temporal stability of the Implicit Association Test-anxiety. Journal of Personality Assessment, 84(1), 82-88.

Fazio, R. H., Sanbonmatsu, D. M., Powell, M. C., & Kardes, F. R. (1986). On the automatic activation of attitudes. Journal of Personality and Social Psychology, 50(2), 229-238.

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

Ferguson, M. J., & Bargh, J. A. (2003). The constructive nature of automatic evaluation. In Musch, J. & Klauer, K. C. (Eds). The Psychology of Evaluation: Affective Processes in Cognition and Emotion (pp. 169-188). Mahwah, NJ: Lawrence Erlbaum Associates

Florack, A., Bless, H., & Piontkowski, U. (2003). When do people accept cultural diversity? Affect as determinant. International Journal of Intercultural Relations, 27, 627-640.

Foroni, F., & Mayr, U. (in press). The power of a story: New, automatic associations from a single reading of a short scenario. Psychonomic Bulletin and Review.

Gehring, W. J., Karpinski, A., & Hilton, J. L. (2003). Thinking about interracial interactions. Nature Neuroscience, 6(12), 1241-1243.

Gemar, M.C., Segal, Z.V., Sagrati, S., & Kennedy, S.J. (2001). Mood-induced changes on the Implicit Association Test in recovered depressed patients. Journal of Abnormal Psychology, 110(2), 282-289.

Gilbert, D. T., & Hixon, J. G. (1991). The trouble of thinking: Activation and application of stereotypic beliefs. Journal of Personality and Social Psychology, 60, 509-517.

Govan, C. L., & Williams, K. D. (2004). Reversing or eliminating IAT effects by changing the affective valence of the stimulus items. Journal of Experimental Social Psychology, 40(3), 357-365.

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. Psychological Review, 102(1), 4-27.

Greenwald, A. G., & Farnham, S. D. (2000). Using the Implicit Association Test to measure self-esteem and self-concept. Journal of Personality and Social Psychology, 79(6), 1022-1038.

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

Greenwald, A. G., & Nosek, B. A. (2001). Health of the Implicit Association Test at age 3. Zeitschrift für

Experimentelle Psychologie, 48, 85-93.

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

Greenwald, A. G., Nosek, B. A., Banaji, M. R., & Klauer, K. C. (in press). Validity of the Salience Asymmetry Interpretation of the IAT: Comment on Rothermund and Wentura (2004). Journal of Experimental Psychology: General.

Hall, G., Mitchell, C., Graham, S., & Lavis, Y. (2003). Acquired equivalence and distinctiveness in human discrimination learning: Evidence for associative mediation. Journal of Experimental Psychology: General, 132(2), 266-276.

Hofmann, W., Gawronski, B., Gschwendner, T., Le, H., & Schmitt, M. (in press). A meta-analysis on the correlation between the Implicit Association Test and explicit self-report measures. Personality and Social Psychology Bulletin.

Hummert, M. L., Garstka, T. A., O'Brien, L. T., Greenwald, A. G., & Mellott, D. S. (2002). Using the implicit association test to measure age differences in implicit social cognitions. Psychology and Aging, 17(3), 482-495.

Jajodia, A., & Earleywine, M. (2003). Measuring alcohol expectancies with the implicit association test. Psychology of Addictive Behaviors, 17(2), 126-133.

Kahneman, D., & Treisman, A. (1984). Changing views of attention and automaticity. In R. Parasuraman & D. R. Davies (Eds.), Varieties of Attention (pp. 29-61). San Diego, CA: Academic Press.

Kahneman, D., & Tversky, A. (1974). Heuristics and Biases.

Karpinski, A., & Hilton, J. L. (2001). Attitudes and the implicit association test. Journal of Personality and Social Psychology, 81, 774-788.

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

Klauer, K. C., & Mierke, J. (2004). Task-set inertia, attitude accessibility, and compatibility-order effects: New evidence for a task-set switching account of the IAT effect. Unpublished manuscript: Rheinische Friedrich-Wilhelms-Universitat Bonn.

Lowery, B. S., Hardin, C. D., & Sinclair, S. (2001). Social influence effects on automatic racial prejudice. Journal of Personality and Social Psychology, 81, 842-855.

Macrae, C. N., Bodenhausen, G. V., Milne, A. B., Thorn, T. M. J., & Castelli, L. (1997). On the activation of social stereotypes: The moderating role of processing objectives. Journal of Experimental Social Psychology, 33, 471-489.

Maison, D., Greenwald, A. G., Bruin, R. (2001). The Implicit Association Test as a measure of implicit consumer attitudes. Polish Psychological Bulletin, 32(1), 61-69.

Marsh, K. L., Johnson, B. T., & Scott-Sheldon, L. A. (2001). Heart versus reason in condom use: Implicit versus explicit attitudinal predictors of sexual behavior. Zeitschrift fuer Experimentelle Psychologie, 48(2), 161-175.

McFarland, S. G., & Crouch, Z. (2002). A cognitive skill confound on the Implicit Association Test. Social Cognition, 20(6), 483-510.

Mierke, J., & Klauer, K. C. (2001). Implicit association measurement with the IAT: Evidence for effects of executive control porcesses. Zeitschrift fuer Experimentelle Psychologie. 48(2), 107-122.

Mierke, J. & Klauer, K. C. (2003). Method-specific variance in the Implicit Association Test. Journal of Personality and Social Psychology, 85(6), 1180-1192.

Mitchell, C. J. (2004). Mere acceptance produces apparent attitude in the Implicit Association Test. Journal of Experimental Social Psychology, 40(3), 366-373.

Mitchell, J. A., Nosek, B. A., & Banaji, M. R. (2003). Contextual variations in implicit evaluation. Journal of Experimental Psychology: General, 132(3), 455-469.

Monteith, M. J., Ashburn-Nardo, L., Voils, C. I., Czopp, A. M. (2002). Putting the brakes on prejudice: On the development and operation of cues for control. Journal of Personality and Social Psychology, 83(5), 1029-1050.

Monteith, M. J., Voils, C. I., & Ashburn-Nardo, L. (2001). Taking a look underground: Detecting, interpreting, and reacting to implicit racial biases. Social Cognition, 19(4), 395-417.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. Psychological Review, 84(3), 231-259.

Nosek, B. A. (2005). Moderators of the relationship between implicit and explicit attitudes. Unpublished manuscript.

Nosek, B. A., & Banaji, M. R. (1997). Initial parameter testing of the Implicit Association Test. Unpublished data.

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

Nosek, B. A., Banaji, M. R., & Greenwald, A. G. (2002). Harvesting intergroup implicit attitudes and beliefs from a demonstration Web site. Group Dynamics, 6(1), 101-115.

Nosek, B. A., Greenwald, A. G., & Banaji, M. R. (2004). Understanding and using the Implicit Association Test: II. Method variables and construct validity. Personality and Social Psychology Bulletin, 31(2), 166-180.

Nosek, B. A., & Smyth, F. L. (2004). A multitrait-multimethod validation of the Implicit Association Test: Implicit and explicit attitudes are related but distinct constructs. Unpublished manuscript.

Nunnally, J. C., & Bernstein, I. H. (1994). Psychometric Theory. New York, NY: McGraw-Hill.

Olson, M. A., & Fazio, R. H. (2003). Relations between implicit measures of prejudice: What are we measuring? Psychological Science, 14(6), 636-639.

Ottaway, S. A., Hayden, D. C., & Oakes, M. A. (2001). Implicit attitudes and racism: Effects of word familiarity and frequency on the implicit association test. Social Cognition, 19(2), 97-144.

Perugini, M., & Leone, L. (2004). Individual differences in moral decision making: Validation of an implicit measure of morality. Unpublished manuscript.

Phelps, E. A., O'Connor, K. J., Cunningham, W. A., Funayama, E. S., Gatenby, J. C., Gore, J. C., & Banaji, M. R. (2000). Performance on indirect measures of race evaluation predicts amygdala activation. Journal of Cognitive Neuroscience, 12(5), 729-738.

Poehlman, T. A., Uhlmann, E., Greenwald, A. G., & Banaji, M. R. (2004). Understanding and Using the Implicit Association Test: III. Meta-analysis of Predictive Validity. Manuscript submitted for publication.

Richeson, J. A., & Ambady, N. (2003). Effects of situational power on automatic racial prejudice. Journal of Experimental Social Psychology, 39(2), 177-183.

Richeson, J. A., Baird, A. A., Gordon, H. L., Heatherton, T. F., Wyland, C. L., Trawalter, S., & Shelton, J. N. (2003). An fMRI investigation of the impact of interracial contact on executive function. Nature Neuroscience, 6(12), 1323-1328.

Richeson, J. A., & Nussbaum, R. J. (2004). The impact of multiculturalism versus color-blindness on racial bias. Journal of Experimental Social Psychology, 40(3), 417-423.

Roediger, H. L. (1990). Implicit memory: Retention without remembering. American Psychologist, 45(9), 1043-1056.

Rothermund, K., & Wentura, D. (2001). Figure-ground asymmetries in the Implicit Association Test (IAT). Zeitschrift fuer Experimentelle Psychologie, 48(2), 94-106.

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

Rudman, L. A., Ashmore, R. D., & Gary, M. L. (2001). “Unlearning” automatic biases: The malleability of implicit prejudice and stereotypes. Journal of Personality and Social Psychology, 81(5), 856-868.

Rudman, L. A., Greenwald, A. G., Mellott, D. S., & Schwartz, J. L. K. (1999). Measuring the automatic components of prejudice: Flexibility and generality of the Implicit Association Test. Social Cognition, 17(4), 437-465.

Rudman, L. A., & Heppen, J. B. (2003). Implicit romantic fantasies and women's interest in personal power: A glass slipper effect? Personality and Social Psychology Bulletin, 29(11), 1357-1370.

Schacter, D. L., Bowers, J., & Booker, J. (1989). Intention, awareness, and implicit memory: The retrieval intentionality criterion. In S. Lewandowsky, J. C. Dunn, & K. Kirsner (Eds.), Implicit memory: Theoretical issues (pp. 47–65).

Schmukle, S. C., & Egloff, B. (in press). A latent state-trait analysis of implicit and explicit personality measures. European Journal of Psychological Assessment.

Schmukle, S. C., & Egloff, B. (2004). Does the Implicit Association Test for assessing anxiety measure trait and state variance? European Journal of Personality, 18, 483-494.

Sherman, S. J., Presson, C. C., Chassin, L., Rose, J. S., & Koch, K. (2003). Implicit and explicit attitudes toward cigarette smoking: The effects of context and motivation. Journal of Social and Clinical Psychology, 22, 13-39.

Simon, H. A. (1955). A behavioral model of rational choice. *Quarterly Journal of Economics*, *69*, 99-118.

Sriram, N., & Greenwald, A. G. (2005). Biases in mean response latency differences. Unpublished Manuscript.

Steffens, M. C. (in press). Is the Implicit Association Test immune to faking? Experimental Psychology.

Steffens, M. C., & Buchner, A. (2003). Implicit Association Test: Separating transsituationally stables and variable components of attitudes toward gay men. Experimental Psychology, 50(1), 33-48.

Steffens, M. C., & Plewe, I. (2001). Items' cross-category associations as a confounding factor in the Implicit Association Test. Zeitschrift für Experimentelle Psychologie, 48, 123-134.

Teachman, B. A., Gregg, A. P., Woody, S. R. (2001). Implicit associations for fear-relevant stimuli among individuals with snake and spider fears. Journal of Abnormal Psychology, 110(2), 226-235.

Teachman, B. A. Woody, S. R. (2003). Automatic processing in spider phobia: Implicit fear associations over the course of treatment. Journal of Abnormal Psychology, 112(1), 100-109.

Teachman, B. A, Gapinski, K. D., Brownell, K. D., Rawlins, M., Jeyaram, S. (2003). Demonstrations of implicit anti-fat bias: The impact of providing causal information and evoking empathy. Health Psychology. 22(1), 68-78.

Teige, S., Schnabel, K., Banse, R., & Asendorpf, J. B. (2004). Assessment of multiple implicit self-concept dimensions using the Extrinsic Affective Simon Task (EAST). Unpublished manuscript.

Wegner, D. M. (2002). The Illusion of Conscious Will. Bradford Books: Cambridge, MA.

Wigboldus, D. (in press). Single-target implicit associations. Experimental Psychology.

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

Wittenbrink, B., Judd, C. M., & Park, B. (1997). Evidence for racial prejudice at the implicit level and its relationship with questionnaire measures. Journal of Personality and Social Psychology, 72(2), 262-274.

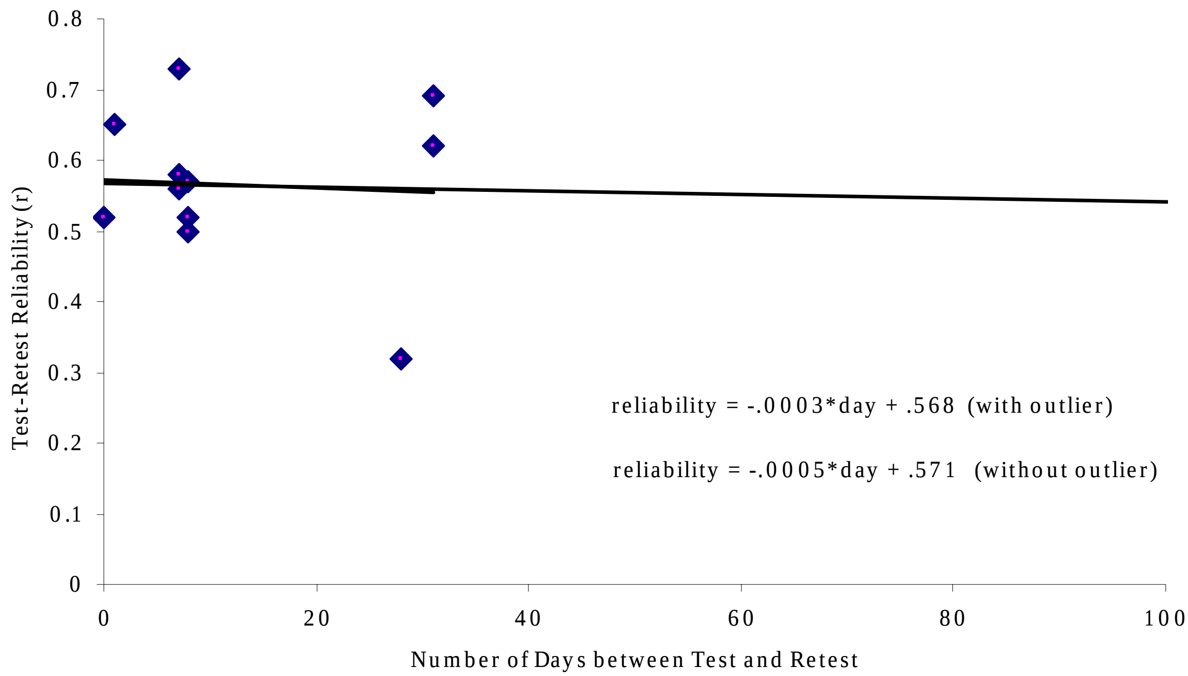
Wittenbrink, B., Judd, C. M., & Park, B. (2001). Spontaneous prejudice in context: Variability in automatically activated attitudes. Journal of Personality and Social Psychology, 81(5), 815-827.

Table 1. Sequence of blocks in the an IAT measuring gender evaluations

Block	No. of trials	Items assigned to left-key response	Items assigned to right-key response
B1	20	Faces of females	Faces of males
B2	20	Pleasant words	Unpleasant words
B3	20	Female faces + Pleasant words	Male faces + Unpleasant words
B4	40	Female faces + Pleasant words	Male faces + Unpleasant words
B5	40	Faces of males	Faces of females
B6	20	Male faces + Pleasant words	Female faces + Unpleasant words
B7	40	Male faces + Pleasant words	Female faces + Unpleasant words

Note. A trial is defined as the time from the onset of a single stimulus to the correct categorization of that stimulus. Trials in which an error is made require the participant to correct the error before proceeding. Blocks B3, B4, B6, and B7 alternate trials presenting a pleasant or unpleasant word with trials presenting a male or female face. If all of the stimuli were from the same modality (e.g., words), then items and labels from one response dimension (men/women) would be presented in a distinct font (color or type) from the items and labels from the other dimension (pleasant/unpleasant). In most IAT applications, the sorting rules in blocks B1, B3, B4 are counterbalanced with B5, B6, B7 between subjects. IAT = Implicit Association Test.

Figure 1. Test-retest reliability for the IAT by number of days between test and retest.



Note: Adapted from Egloff et al., 2005 with additional studies included. Data points represent Banse et al., 2001; Bosson et al., 2000; Cunningham, et al., 2001; Dasgupta & Greenwald, 2001; Egloff et al., 2005; Egloff & Schmukle, 2002; Greenwald & Farnham, 2000; Schmukle & Egloff, 2004; and, Steffens & Buchner, 2003. Outlier data-point appears at 379 days with an r of .47. IAT = Implicit Association Test.