Psychological Inquiry
2003, Vol. 14, No. 3&4, 238–243

The First Ontological Challenge to the IAT: Attitude or Mere Familiarity?

Nilanjana Dasgupta
Department of Psychology
University of Massachusetts—Amherst

Anthony G. Greenwald
Department of Psychology
University of Washington

Mahzarin R. Banaji
Department of Psychology
Harvard University

Somebody once asked "Why is it that when people say 'that's a good question' they never have a good answer?" In response to the query of how we came to do this work, "good question" was indeed our own response, and as such we cannot promise to have a good answer. In spite of the irony that this exercise poses for us, who insist on a healthy distrust of introspective analysis, in this article we hope to communicate the many pleasures of our collaborative effort, the degree to which we are indebted to our critics, and the recognition that the larger understanding of implicit social cognition involves many others who constitute an integral part of this discovery.

A Brief History

The origins of the work chosen for this issue lie in the development of the Implicit Association Test (IAT). Since the late 1980s Banaji and Greenwald had been testing various effects that were captured by the umbrella term "implicit social cognition." The first of these demonstrated the usage of an implicitly activated gender stereotype linking men with fame (Banaji & Greenwald, 1995), and a review article that reinterpreted existing findings as evidence for implicit social cognition in the domains of attitudes, stereotyping and self-esteem (Greenwald & Banaji, 1995).

A feature of Greenwald's work was humorously captured by Bob Abelson who confided to Banaji that "When everybody gets an effect, Tony gets no effect; when everybody gets a large effect, Tony gets a small one." Having been trained by this master of nullish effects, Banaji was unperturbed by experiments that produced small but reliable effects using standard priming techniques to measure the strength of stereotypes (Banaji & Hardin, 1996; Blair & Banaji, 1996) and in fact gave a lecture entitled "The importance of an 8 millisecond effect." But Greenwald had had it with a life of small to null effects and, with a vengeance, sought to build a tool that would capture implicit social cognition and consistently produce robust and easily replicable results. He worked on developing a task—the IAT—that appeared in a grant proposal submitted in collaboration with Banaji. Here is what the NIH review committee saw by way of the first short description of this task in a proposal of approximately 40 experiments.

> Experiment 3.8: *Measurement of implicit attitude (B: Rapid classification method).* The same materials as Expt. 3.7 are used, but without priming. Instead, two categories of words are assigned to each of two response keys. Subjects are asked to rapidly press (say) the right key whenever the stimulus word is *either* female-associated or pleasant in meaning, and the left key for words either male-associated or unpleasant in meaning. Through the course of a session, blocks of trials with the four combinations of category pairings and key assignments are intermixed. Because early trials in any block should be heavily contaminated by response assignments from preceding blocks, the data would be regression analyzed with multiple predictors including block number, trial number within block, and interactions of these with other design factors. The aim of the regression analysis is to model and remove effects due to the switching assignments of category pairs to keys. The measure of implicit attitude (abstracted from the regression analysis) is the difference between latency with pleasant/male pairing versus pleasant/female pairing. To the extent that responding is faster with pleasant/female than with pleasant/male pairing, the latency-difference measure indicates greater positivity of the implicit attitude associated with female.

Greenwald conducted the first IAT on himself, then on his collaborators. Each was stunned by the difficulty in making equally rapid associations between certain

concept+attribute pairs (e.g., Black+good/White+bad) compared to others (Black+bad/White+good). Our initial discomfort with the result was immediately replaced by the recognition that the task captured something important about the automatic aspects of social cognition and its unique properties as a consciousness raising device. It also produced effects that reminded us of a fast food place in Columbus, OH, called "Burritos As Big As Your Head." Long before the IAT was published (Greenwald, McGhee, & Schwartz, 1998), the decision was made to give it away to any group of researchers interested in exploring it, and several colleagues and students responded with enthusiastic attempts to figure out its ontology and thereby its meaning.

From the moment the Implicit Association Test hit the conference circuit, the scholarly press, and the Internet, reactions came pouring in from researchers and lay people alike. Not surprisingly, the specific task that most people were reacting to was not the IAT measuring automatic attitudes toward flowers and insects (which they agreed was probably measuring their favorable attitude toward flowers compared to insects), but rather the one measuring attitudes toward racial groups. These reactions, both positive and negative, were remarkably passionate and remarkably uniform irrespective of whether they were from other scientists or lay people. Even those not familiar with the term "counterbalancing" wrote about the possible explanation that their responses during the task may have varied as a function of the order in which they received the two conditions that make up each IAT.

Among the most persistent of questions involved a particular alternative interpretation of why so many of the participants, especially non-Black participants, found it easier to associate the category White American with positive attributes and African American with negative attributes than vice versa. Put differently, at issue was the following question: were the speeded reactions captured by the IAT driven by people's underlying automatic racial *attitudes* or by some other confounding variable? This essay traces the development of an empirical article that sought to understand how best to interpret the response latency data captured by the race IAT and similar other reaction time tasks claiming to measure intergroup prejudice (see Dasgupta, McGhee, Greenwald, & Banaji, 2001).

Behind the Scenes

When the first paper debuting the IAT as a new measure of automatic attitudes was being written in 1997 (Greenwald et al., 1998), one of us (Dasgupta) was metamorphosing from a graduate student into a postdoctoral fellow during the long drive from the east coast (New Haven, CT) to the west coast (Seattle, WA). Besides having one family member in common, these two social psychology labs shared other, more important, commonalities. Members of both labs on both coasts were noticeably galvanized by our subjective experience of taking this test of stroop-like difficulty. Because of our deeply held beliefs about equal treatment, our own automatic responses were difficult to explain away. Now we could no longer talk about "those others" who held negative attitudes toward disadvantaged groups. We were repeatedly made aware that we were responding noticeably faster, and making fewer errors, when we had to pair White (or young) with pleasant stimuli and Black (or old) with unpleasant stimuli, and that we could not will our automatic responses to become aligned with our conscious beliefs about race or age. There is no question that among the appealing aspects of the IAT was the ability of the experimenters to be subjects in their own experiments. E-mail between the labs was buzzing with variations being tried and reports of yet another confession of one's automatic attitude. If we were at all in doubt about the fascination the task held for us, it was immediately confirmed by the disbelief and resistance we received from our family, friends, and peers whenever IAT demonstrations and data were presented at conferences, lectures, and in the media.

Both the magnitude of attention this research attracted and the specific content of the questions posed (about internal validity, construct validity, and predictive validity) animated lab group meetings and motivated us to develop experiments aimed at testing each of the methodological and theoretical questions that were sent our way. There's no doubt that these questions, comments, and criticisms accelerated our collective productivity—making our research on automatic prejudice and stereotyping move forward faster than it would have otherwise.

Initially, both labs in Seattle and New Haven focused on testing the parameters of the task itself by varying stimuli, labels, timing, the number of trials, and so on. We also examined order effects and experimented with the computation of IAT scores. This continues to remain an important component of our work even today (see Greenwald, Nosek, & Banaji, 2003). Among the many questions students were pursuing in Seattle and New Haven, here is a sampling. If people are explicitly told that the race IAT measures their automatic attitudes about racial groups, can they deliberately correct their responses to exhibit less bias? What if they are given a specific strategy with which to avoid bias—does that prevent the expression of automatic prejudice? Can automatic bias against outgroups be reduced without relying on people's motivation and control over their responses? How stable are people's automatic attitudes across time? Does the relationship between self-reported attitudes and IAT-assessed automatic attitudes vary in magnitude depending on the order in which these tasks are completed? And the first

clear alternative explanation: Can the automatic bias captured by the IAT be explained by the subjective lack of familiarity with one group (Black) compared to the other (White) or by the objectively lower frequency of occurrence of the category Black versus White in mainstream American culture?

The new postdoc chose to pursue the last question for two reasons. First, given the frequency with which this question was being raised, it was clear that this basic issue of construct validity needed to be addressed before other more complex issues could be tackled. The question was ripe for testing and it was unclear whether or not the entire effect could be explained away by mere familiarity. Second, the issue of familiarity and frequency was linked to a larger, older, and more interesting question in social psychology: what is the relationship between people's familiarity with social groups and their attitudes toward them? At a macro level, we know that increased familiarity with (e.g., interpersonal contact with) outgroup members, increases liking for those outgroups under some conditions (Amir, 1969; Brewer, 1996; Pettigrew, 1998; Pettigrew & Tropp, 2000). But what about the micro level? Do people's evaluations of racial groups fluctuate as a function of their familiarity with specific instantiations of those groups used in any given experiment? This was a particularly viable criticism of the race IAT at the time given that until 1998 racial groups were always represented with stereotypically Black (Latoya and Tyrone) and White (Wendy and Brandon) first names. To the extent that there are fewer Latoyas and Tyrones in the American population than there are Wendys and Brandons, the difference in the objective frequency with which these types of names occur in the population is likely to elicit differential feelings of familiarity in perceivers. In other words, because of their low frequency, racially identifiable Black names are likely to be less familiar to perceivers than racially identifiable White names. To the extent that greater familiarity leads to more liking (Zajonc, 1968), White names may accrue more positivity than Black names making it easier for people to associate White with good and Black with bad than vice versa.

Frequency and familiarity aside, the use of stereotypic Black names may also confound race and class. Some evidence suggests that racially identifiable Black first names such as the ones used by Greenwald et al. (1998) may be more common within lower socioeconomic Black communities than middle- or upper-middle class Black communities (Lieberson & Bell, 1992). If participants in the Greenwald et al. (1998) study were sensitive to class differences between the Black and White names, it leaves open the possibility that the data may be better interpreted as revealing people's automatic attitudes toward social class rather than toward race. An additional trouble with names was that many of the names identified as White were possessed by African Americans and participants who knew such individuals had particular difficulty with the classification.

Multiple Ways to Slay a Dragon

We designed three tests to rule out the potential effect of name familiarity on automatic race bias. First, we measured participants' familiarity with Black and White names before measuring their automatic racial attitudes. In order to minimize method variance and the likelihood of socially desirable responding, we created a response latency measure of familiarity instead of simply asking for a self-report of participants' familiarity with each name. Specifically, we asked participants to discriminate real names from pseudonames as quickly and accurately as possible; sometimes the real names in the name recognition task were White names and at other times they were Black names. Our logic was that the speed with which participants differentiated White names from pseudonames compared to Black names from pseudonames could be taken as an indicator of their relative familiarity with White compared to Black first names. Drawing inspiration from *Star Trek* (Roddenberry, 1966–1969; all of us being staunch followers of the show) we created a number of fake Klingon-type names like Nekar, Bralla, Arton, Anadri to serve as pseudonames. The relationship between name familiarity and automatic race bias was examined using a statistical regression technique developed by Greenwald, Klinger, and Schuh (1995) to determine whether the race bias effect (i.e., the IAT effect) remained significantly different from zero even when both types of names were perceived to be equally familiar.

We also approached familiarity from a different angle: by replacing first names with pictures of unfamiliar Black and White individuals downloaded from a college yearbook. If differences in name familiarity were responsible for producing automatic race bias, then the effect should vanish when pictures instead of names were used. Likewise, if class differences inherent in the names were responsible for automatic race bias, then again the effect should evaporate upon using pictures. The results were clear. Both name and picture IATs revealed strong preference for White Americans relative to African Americans, and the effect remained robust even after name familiarity was statistically controlled.

Stimulus familiarity was tackled from yet another angle in a follow-up study in which we manipulated the objective frequency of Black and White names across four IATs using name frequency counts from the Internet, validated against the 1990 U.S. census data (U.S. Department of Justice, 1998). We found that, contrary to the stimulus familiarity explana-

tion, a larger White preference effect emerged when Black and White names were matched in frequency than when popular White names were contrasted against rare Black names in the IAT. Although we regard this study to be quite important in ruling out the familiarity explanation, the reviews led to this study being dropped from final publication in the *Journal of Experimental Social Psychology* (*JESP*). For those who are interested, a report of that study is available in the resources section at www.people.fas.harvard.edu/~banaji.

With a Little Help From Our Friends

It should be made quite clear that we were not alone in tackling the familiarity problem. While our studies were being conducted in Seattle, Scott Ottaway and his colleagues (Ottaway, Hayden, & Oakes, 2001), Laurie Rudman and her colleagues (Rudman, Greenwald, Mellott, & Schwartz, 1999), and Brian Nosek and his colleagues (Nosek, Banaji, & Greenwald, 2002) were conducting other studies to test the role of stimulus familiarity on automatic attitudes measured by various IATs. Ottaway et al. (2001) found that even when Hispanic, Black, and White names were equated in terms of objective frequency and subjective familiarity, people still nonconsciously favored White Americans over Black and Hispanic Americans.

Extending the test of stimulus familiarity beyond race, Rudman et al. (1999) compared people's automatic attitudes toward Christians versus Jews, young versus old people, and American versus Russian leaders. They found that participants expressed pro-Christian, pro-young, and pro-American attitudes even when stimulus names representing the target groups were selected to be equally familiar or name familiarity was statistically controlled. In fact, in the case of pro-U.S. sentiments, American participants were found to favor obscure American presidents over familiar Russian presidents, suggesting that preference for one's ingroup was clearly overriding perceivers' unfamiliarity with specific instantiations of the ingroup.

Reaching even further to test the role of familiarity in driving automatic attitudes toward academic disciplines, Nosek et al. (2002) showed that people's familiarity with words and symbols related to mathematics did not always result in automatic liking for that discipline as measured by the IAT. Women favored unfamiliar stimuli (specifically, little known geographical locations) over familiar mathematical concepts whereas men favored mathematics over unfamiliar places. It was the collective impact of these multiple studies that decisively laid the familiarity explanation to rest. That it ceased to be quite the issue it had been, became evident when familiarity dropped off the list of criticisms after the publication of this paper. Now, we and others could move on to other tests of validity and generalizability.

Among the other studies that had an impact similar to this one is the study by Phelps et al. (2001). Just as with the question of familiarity, another issue routinely raised about the IAT was this: the IAT may measure something but that something is not an attitude (see Banaji, 2001, for a discussion). That is, the task detects something cold and cognitive not anything warm and affective. It is our guess that dozens of behavioral studies may not have been able to answer this question to the satisfaction of most critics. However, by showing that the magnitude of preference for White versus Black faces on the IAT was related to the differential activation of the amygdala in response to the same faces put that concern to rest as well. The finding that the amygdala, a sub-cortical structure known to be involved in emotional learning and memory, was associated with IAT responses lent some credence to the assertion that the IAT was capturing something warm and affect-laden.

Coda

Although both the *JESP* study on familiarity and the cognitive neuroscience study on the sub-cortical correlates of attitudes have answered specific questions about the construct validity of the IAT, they leave other related questions unanswered. Much work still needs to be done to address these issues. Speaking only of familiarity, the research described here addresses the issue of familiarity at the micro level: that is, does lack of familiarity with specific stimuli representing outgroups (in this case, names) in various IATs produce automatic bias against those groups?

However, this research does not address the issue of intergroup familiarity at the macro level. For an answer to that question, we go back to an older literature on the contact hypothesis that found lack of experience with particular social groups to indeed be related to prejudice (Allport, 1954; Amir, 1969; Aronson & Bridgeman, 1979; Cook, 1969; Pettigrew, 1997). When experience with outgroups is enhanced via interpersonal contact or experience with outgroup members, attitudes toward those groups become more positive provided specific conditions are met (Brewer, 1996; Herek & Capitanio, 1996; Hewstone, 2000; Pettigrew & Tropp, 2000). No doubt lack of experience with outgroups is partly responsible for fanning the flames of both implicit and explicit prejudice. At the same time, prejudice may also help maintain inexperience with particular groups. Individuals who harbor prejudice against any given group (implicitly or explicitly) are likely to avoid situations that would increase their knowledge of that group and its culture, but instead to seek out situations populated

by like-minded ingroup members, both of which may foster greater prejudice. If, however, their social environment provided frequent exposure to outgroup members, especially clearly admirable ones, automatic prejudicial attitudes may recede (Dasgupta & Greenwald, 2001; Towles-Schwen & Fazio, 2001).

We believe that research explicating the relationship between intergroup familiarity and implicit and explicit prejudice is likely to be an important topic in the 21st century as globalization and immigration continue to change the demographics of the United States, and indeed many other countries around the world. In order for social groups to co-exist, and to do so justly, we need to better understand the social processes, like intergroup contact and experience, that may, under the right conditions, attenuate prejudice and stereotyping. The discussion here is primarily meant to point out that although our paper had impact in ruling out stimulus familiarity as an alternative explanation of automatic intergroup bias, we agree that familiarity or experience with outgroups is likely to be part and parcel of intergroup attitudes. Nevertheless, we are also persuaded that equalizing familiarity may not fully erase biases in intergroup attitudes. A question posed (and deftly answered) by Steve Pinker at a talk one of us gave at MIT is perhaps instructive here. His question concerned whether lack of familiarity was linked to a particular attribute of some social categories—i.e., that of being a marked category. African Americans are a marked category, just as Jews and Asians may be, because of their position in American society. Perhaps this accounted for the ease with which some stimuli were paired in the IAT—i.e., marked groups with negative attributes. He answered his own question upon hearing the finding that the basic IAT attitude effect does not operate in the context of gender—although women constitute a marked category, they do not evoke negativity. Gender may also address the familiarity question to some extent. Women coexist with men to a greater extent than Black Americans do with White Americans. If lack of familiarity alone was responsible for outgroup bias, then women's familiarity with men ought to attenuate their automatic bias against men. Yet, women's attitude toward men as measured by IATs is substantially more negative than their attitude toward their ingroup. Studies of other such groups that cohabitate and yet show attitude differences will be instructive in this regard.

Susan Fiske points out in her essay in this issue that the articles she selected share certain properties, one of which is that they rub people the wrong way. We are glad that this work rubbed some very intelligent people the wrong way and that it continues to do so. Without it, our lives would have been far more simple, but far more boring.

Note

Nilanjana Dasgupta, Department of Psychology, Tobin Hall, University of Massachusetts-Amherst, Amherst, MA 01003. E-mail: dasupta@psych.umass.edu

References

Allport, G. W. (1954). *The nature of prejudice.* Reading, MA: Addison-Wesley.

Amir, Y. (1969). Contact hypothesis in ethnic relations. *Psychological Bulletin, 71,* 319–342.

Aronson, E., & Bridgeman, D. (1979). Jigsaw groups and the desegregated classroom: In pursuit of common goals. *Personality and Social Psychology Bulletin, 5,* 438–446.

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

Banaji, M. R., & Hardin, C. D. (1996). Automatic stereotyping**.** *Psychological Science, 7,* 136–141.

Banaji, M. R., & Greenwald, A. G. (1995). Implicit gender stereotyping in judgments of fame. *Journal of Personality and Social Psychology, 68,* 181–198.

Blair, I. V., & Banaji, M. R. (1996). Automatic and controlled processes in stereotype priming. *Journal of Personality & Social Psychology, 70,* 1142–1163.

Brewer, M. B. (1996). When contact is not enough: Social identity and intergroup cooperation. *International Journal of Intercultural Relations, 20,* 291–303.

Cook, S. W. (1969). Motives in a conceptual analysis of attitude-related behavior. *Nebraska Symposium on Motivation, 17,* 179–231.

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,* 800–814.

Dasgupta, N., McGhee, D. E., Greenwald, A. G., & Banaji, M. R. (2001). Automatic preference for White Americans: Eliminating the familiarity explanation. *Journal of Experimental Social Psychology, 36,* 316–328.

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

Greenwald, A. G., Klinger, M. R., & Schuh, E. S. (1995). Activation by marginally perceptible ("subliminal") stimuli: Dissociation of unconscious from conscious cognition. *Journal of Experimental Psychology: General, 124*, 22–42.

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,* 1464–1480.

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

Herek, G. M., & Capitanio, J. P. (1996). "Some of my best friends": Intergroup contact, concealable stigma, and heterosexuals' attitudes toward gay men and lesbians. *Personality and Social Psychology Bulletin, 22,* 412–424.

Hewstone, M. (2000). Contact and categorization: Social psychological interventions to change intergroup relations. In C. Stangor (Ed.), *Stereotypes and prejudice: Essential readings* (pp. 394–418). Philadelphia: Psychology Press.

Lieberson, S., & Bell, E. O. (1992). Children's first names: An empirical study of social taste. *American Journal of Sociology, 98,* 511–554.

Nosek, B. A., Banaji, M. R., & Greenwald, A. G. (2002). Math = male, me = female, therefore math ≠ me. *Journal of Personality and Social Psychology, 83,* 44–59.

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*, 97–144.

Pettigrew, T. F. (1997). Generalized intergroup contact effects on prejudice. *Personality & Social Psychology Bulletin, 23,* 173–185.

Pettigrew, T. F. (1998). Intergroup contact theory. *Annual Review of Psychology, 49*, 65–85.

Pettigrew, T. F., & Tropp, L. R. (2000). Does intergroup contact reduce prejudice: Recent meta-analytic findings. In S. Oskamp (Ed.), *Reducing prejudice and discrimination* (pp. 93–114). Mahwah, NJ: Erlbaum.

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

Roddenberry, E. W. (1966–1969). *Star Trek* (Television series). New York: NBC.

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

Towles-Schwen, T., & Fazio, R. H. (2001). On the origins of racial attitudes: Correlates of childhood experiences. *Personality and Social Psychology Bulletin, 27,*162–175.

U. S. Department of Justice. (1998). *Database of first names from the 1990 U. S. Census* [Online]. Available at http://www.census.gov/geneology/names/ [1998, January 7].

Zajonc, R. B. (1968). Attitudinal effects of mere exposure. *Journal of Personality and Social Psychology, Monographs, 9*(2, Pt. 2).