Exhibit 5
Declaration and Report
of
Anthony G. Greenwald

UNIVERSITY LEGAL ASSISTANCE
Larry A. Weiser, Attorney at Law
Jacob White, Legal Intern
Kristine Olmstead, Legal Intern
721 North Cincinnati Street
P.O. Box 3528
Spokane, Washington 99220
(509)323-5791 Telephone
(509)323-5805 Facsimile

Attorneys for Plaintiffs

UNITED STATES DISTRICT COURT
EASTERN DISTRICT OF WASHINGTON

MUHAMMAD SHABAZZ FARRAKHAN et. al., Plaintiffs, v. CHRISTINE O. GREGOIRE, et. al., Defendants.	NO: CV-96-076-RHW DECLARATION OF EXPERT WITNESS ANTHONY GREENWALD

I, Anthony Greenwald, declare and state as follows:

1. I am over 18 years of age and am competent to testify to the matters stated herein.

2. Attached herein is a true and correct copy of my expert report and my curriculum vitae.

3. Pursuant to FRCP 26(2)(B) I have attached a written report which I have signed and prepared. This document is also a complete statement of all opinions which I will express and the basis and reasons therefore and contains the data or other information considered in forming my opinions. See attached *Expert Report* and the attached article *Behavioral Realism and Implicit Social Cognition.*

4. I have attached a Curriculum Vitae which contains a list of all my publications and qualifications as an expert in the above stated matter. See attached *Curriculum Vitae.*

5. The plaintiffs are paying me $150 per hour for work not related to deposition and court testimony, and $250 per hour for deposition and court testimony for compensation.

6. I have testified as an expert witness in BOB BOGLE vs. GENERAL MOTORS CORPORATION; UNITED STATES DISTRICT COURT, WESTERN DISTRICT OF WASHINGTON, NO. C98-1236C.

I declare under penalty of perjury under the laws of the State of Washington and that the forgoing is true and correct and of my own knowledge.

Signed at Seattle, Washington on this 2nd day of November, 2005.

A. Greenwald

ANTHONY GREENWALD, Declarant

UNIVERSITY LEGAL ASSISTANCE
721 North Cincinnati Street - PO Box 3528
Spokane, WA 99220-3528
(509)323-5791 Telephone
(509)323-5805 Facsimile
(509) 323-3796 TTY

UNIVERSITY LEGAL ASSISTANCE
721 North Cincinnati Street - PO Box 3528
Spokane, WA 99220-3528
(509)323-5791 Telephone
(509)323-5805 Facsimile
(509) 323-3796 TTY

Prepublication draft of article in a special symposium of the *California Law Review*, "Behavioral Realism and Implicit Social Cognition"

Implicit Bias: Scientific Foundations

Anthony G. Greenwald & Linda Hamilton Krieger

Summary

Sections 1–3 of this article review the scientific developments of the past 25 years that have established the foundation for understanding implicit biases. *Implicit biases* are attitudes and stereotypes that operate largely outside of awareness, yet are significant in determining judgment and behavior in relation to stigmatized and disadvantaged groups.

The *Implicit Association Test* (IAT — Section 4) is a research procedure that has greatly accelerated investigations of implicit bias in the last 10 years. Prior investigations of bias have relied mostly on survey (question-asking) methods, which have the two disadvantages of (a) allowing people to choose answers that disguise their actual beliefs and (b) not allowing access to beliefs that escape conscious awareness. The IAT assesses implicit bias without asking questions, instead relying on speeded classification performances that are facilitated or impaired by knowledge that need not be accessible to conscious awareness.

The relation of IAT measures to social judgments and behavior has been investigated in numerous studies, 61 of which were recently summarized in a review that statistically characterizes the aggregate import of these findings. Among the important findings of this review is that IAT measures predict behavior and judgment in

relation to stigmatized groups more effectively than do survey-type measures (Section 5).

The biases revealed by IAT measures are remarkably pervasive. Implicit biases toward specific stigmatized groups often appear at statistically noticeable levels in more than 2/3 of respondents. Pervasiveness of implicit biases is also indicated by their appearance at this high level in virtually every demographic sub-segment of the American population that has been examined among the very large samples of respondents who have taken IATs on demonstration and research web sites (Section 6). *Explicit biases* revealed by survey measures are not nearly so pervasive. It remains a challenge to understand both why implicit biases are so much more pervasive than explicit biases and how implicit biases are formed (Section 7).

Most researchers who have investigated both implicit and explicit bias have concluded that these represent psychologically different forms of bias (Section 8). Implicit biases are seen as being involved in subtle or spontaneous forms of social behavior that can be significant in determining the quality and outcome of social interactions (Section 9). By contrast, explicit biases are conceived as being more involved in deliberate discriminatory behavior that occurs with conscious awareness. Furthermore, conscious awareness may have the potential to override discriminatory effects of implicit bias, when one is aware of possessing implicit bias and of its potential to produce unintended discrimination (Section 10).

Several research studies have established that implicit biases can be reduced by interventions such as exposure to admirable and attractive members of stigmatized

groups. However, such interventions have not yet shown to be capable of producing durable reductions in implicit bias (Section 11).

Two established facts are persuasive in favor of the conclusion (Section 12) that implicit bias is a source of racially disparate outcomes for which race-neutral causes cannot be identified. First is the known pervasiveness of implicit biases that operate outside of awareness and which often deviate from the same person's explicit beliefs. Second is the established relation of implicit bias to behavior that can have discriminatory impact.

Postscript. The study conducted by Katherine Beckett, reported in her May 3, 2004 manuscript, "Race and Drug Law Enforcement in Seattle", illustrates Section 12's point. Beckett identified several race-neutral factors that might have been causally responsible for observed substantial race disparities in drug arrests in Seattle. She identified effective methods of empirically evaluating these potential causes, finding that — singly or in combination — these factors could not explain the observed racial disparity. With evidence favoring racially neutral causes thus lacking, it is plausible, even probable, that the observed racial disparity can be explained by explicit and/or implicit bias. If explicit bias measures had been available to show that the population of arresting police officers was free of such bias, it would remain highly plausible that the observed disparities were due to implicit bias.

Draft of article in press for a special symposium of the *California Law Review*,
"Behavioral Realism and Implicit Social Cognition"
(Footnotes in this draft are in Psychology Journal mode; yet to be converted to Law Review mode)

Implicit Bias: Scientific Foundations

Anthony G. Greenwald[1] & Linda Hamilton Krieger[2]

The assumption of conscious control of human behavior has taken a theoretical battering in recent years. Although this assault in some ways resembles the previous century's Freudian revolution, there are important differences between the two. Freud's views of unconscious mechanism were embedded in a theory that never achieved conclusive support among scientists, despite many, many empirical theory-testing efforts by researchers in the middle third of the 20th century.[3] Consequently, Freud's psychoanalytic theory of unconscious mind and unconscious mental processes has been abandoned by most psychological scientists.

Theoretical conceptions of conscious control over human behavior were strongly re-established in the last third of the 20th century, but this dominance has been crumbling during the past two decades. Unlike the Freudian revolution, however, the new science of unconscious mental process is not the product of a single brilliant theoretical mind.

[1] Professor of Psychology, University of Washington
[2] Professor of Law, University of California, Berkeley (Boalt Hall)
[3] See reviews by Erdelyi (1985), Erdelyi and Goldberg (1979), Greenwald (1992), Kihlstrom (1990), Shevrin and Dickman (1980)

Rather, it is being constructed from an evolving, accumulating body of reproducible empirical findings.[4]

This article aims to provide readers with an understanding of *implicit bias* -- one aspect of the new science of unconscious mental process that has substantial potential bearing on discrimination law. Theories of implicit bias are defined in part by contrast with the "naïve"[5] psychological view of social behavior, which views human actors as being guided by their *explicit* beliefs and conscious intentions. A belief is considered "explicit" if it is consciously believed or endorsed. An intention is considered "conscious" to the extent that the actor is aware of taking an action for a particular reason. In either case, the actor would be *capable* of asserting the belief or identifying the intention, even if *unwilling* to do so.[6]

[4] The early stages of this modern revolution were reviewed by Greenwald (1992). Nisbett and Wilson's (1977) exposé of the inadequacies of introspective explanations of behavior was a noticeable starting point of the modern revolution, leading to widespread understanding that self-report measures of conscious mental process used widely in psychological research were of highly suspect validity. Wegner's (2002) and Bargh's (Bargh et al., 2001) more recent works reveal the frequency with which seemingly ordinary voluntary actions are controlled in ways that evade conscious scrutiny, further undermining the idea that a conscious mind is the effective governor of most human behavior.

[5]The term "naive psychology" refers to laypersons' intuitions about determinants and consequences of human, and especially their own, thought and behavior. Modern treatments were largely inspired by Fritz Heider's (1958) book, *The psychology of interpersonal relations*, which initiated systematic investigation of how laypersons' intuitions differ from scientific understanding.

[6] Methodological investigations by social psychologists in the 1960s revealed social influences operating within research and interview settings that would lead people to describe their explicit beliefs inaccurately in experimental studies (Orne, 1962; Rosenberg, 1969; Weber & Cook, 1972). Work inspired by Festinger's (1957) cognitive dissonance theory initiated modern interest in understanding inability to identify the causes of one's own thought and behavior. Nisbett and Wilson's (1977) article effectively summarized the humbling implications of the ensuing two decades of research.

1. Implicit Cognition

Many different types of mental process have been shown to function implicitly, outside of conscious attentional focus. These include implicit memory,[7] implicit perception,[8] implicit attitudes,[9] implicit stereotypes,[10] implicit self-esteem,[11] and implicit self-concept.[12,13] The meaning of "implicit" in these usages is technical, but still reasonably close to its everyday meaning. For example, implicit memory[14] is memory that is revealed by some performance that could not have occurred in the absence of some previous experience leaving a memory record even though that experience cannot be voluntarily ("explicitly") retrieved. The memory is said to be expressed implicitly, but not explicitly, in the performance. An example: On Day 1 of an experiment[15] subjects were asked to pronounce each of a long list of people's names. Some of these names were recognizably famous, and others were not. On Day 2, these same subjects judged whether each of another long list of names was famous or not. Half of Day 2's non-famous names were repeated from Day 1. The result: More of the *repeated* non-famous names than the novel ones were judged famous. These "false fame" judgments amounted to an implicit memory effect, with the following interpretation: The names acquired some familiarity from Day 1's pronunciation even though, by Day 2, the Day 1 exposure was not likely to be consciously remembered by

[7] Schacter, 1987; Jacoby and Dallas (1981)
[8] Kihlstrom, Barnhardt, and Tataryn (1992)
[9] Greenwald, McGhee, and Schwartz.(1998)
[10] Rudman, Greenwald, and McGhee,(2001)
[11] Greenwald and Farnham (2000)
[12] Rudman et al. (2001)
[13] An overview of implicit social cognition, which encompasses the phenomena of implicit attitudes, stereotypes, self-esteem, and self-concept, is available in Greenwald and Banaji (1995); see also Greenwald et al. (2002).

subjects. This perhaps-vague feeling of familiarity for repeated names was sometimes mis-attributed to fame, leading to greater false judgments of fame for the repeated than the non-repeated names. Subjects are presumably going through a mental process something like: "This name seems familiar. How could it be familiar? Perhaps it's famous." For names that subjects could explicitly remember as having been seen and pronounced on Day 1, subjects would correctly understand why the name seemed familiar and would therefore not mistakenly attribute the familiarity to fame.

Development of any experimentally useful implicit measure requires that the researcher understand the causes of a judgment or action for which the subject is unlikely to discern the cause. The 2-day study just described requires the researcher to understand that pronouncing a name could produce a feeling of familiarity, which could in turn be mis-attributed to fame. Another example requires that the researcher understand causal influences affecting a loudness judgment. The researcher might know, first, that hearing and understanding a word is facilitated after having recently heard it being pronounced and, second, if a word presented in noisy listening conditions is understood it will be judged to have been spoken more loudly than one that can't be understood. Thus, at constant loudness, words that have been made easier to understand by having been heard recently will be judged louder than ones not recently heard.[16] This finding deserves to be called a memory effect because the loudness judgment shows that a trace (i.e., a memory) of the prior exposure exists. It is an

[14]Schacter (1987)
[15]Jacoby, Kelly, Brown, and Jasechko (1989)
[16] Jacoby, Allan, Collins, and Larwill (1988)

implicit effect because the subject remains unaware of the influence of the recent exposure on the loudness judgment.

2. Implicit Attitudes and Implicit Stereotypes

Implicit memory research conducted in the 1980s led researchers to develop measures of other implicit mental phenomena. Two of these — *implicit attitudes* and *implicit stereotypes* — are especially relevant to bias and discrimination.

Implicit Attitudes

Social psychologists define an *attitude* as an evaluative disposition, meaning the tendency to like or dislike, or to act favorably or unfavorably toward, someone or something. Explicit expressions of attitudes occur frequently, whenever we say we like or dislike someone or something. A person's statement that they like a particular presidential candidate provides a ready example. In addition to statements, attitudes can be expressed through favorable or unfavorable *action* — for example, by voting for or against a particular presidential candidate. If the voter understands that the favorable vote results from favorable beliefs about the candidate, the vote can be considered an *explicit attitude expression*.

In other situations, a vote might function as an *implicit attitude indicator*. An implicit attitude is performance of action that indicates favorability or unfavorability toward some object, but is not understood by the actor as expressing an attitude.[17] Consider, for

[17]Greenwald and Banaji (1995, p. 8) define implicit attitude as follows: "Implicit attitudes are introspectively unidentified (or inaccurately identified) traces of past experience that mediate favorable or

example, a situation in which someone votes for a particular candidate even though they know nothing about the candidate other than his or her name. One of the things that might influence a person to vote for this candidate is that the candidate's name shares one or more initial letters with the voter's name. In such a case, the vote could be understood, at least in part, as an implicit expression of the voter's favorable attitude toward self.[18]

As a second example, consider how people go about forming impressions of a liked or disliked candidate's spouse, child, or sibling. Someone who knows nothing about this relative other than the fact of relation to the candidate may feel favorably or unfavorably disposed to the relative. Not too surprisingly, this favorable or unfavorable attitude toward the relative is likely to match the attitude toward the candidate. Evaluation of the unknown relative therefore acts as an *implicit indicator* of attitude toward the candidate. In this example, the "implicit" designation indicates that the attitude expressed toward the relative is determined by the attitude toward the candidate, even though it may be experienced as an independent attitude.

Implicit attitudes are of greatest interest when implicit and explicit attitudes toward the same attitude object differ. These deviations are referred to as *dissociations* between implicit and explicit attitudes. Dissociations are commonly observed in relation to attitudes toward stigmatized groups, including ones defined by race, age, ethnicity,

---

unfavorable feeling, thought, or action toward social objects"; a fuller treatment of implicit attitude phenomena can be found in that article.

[18] see Greenwald and Banaji (1995, p. 12)

disability, and sexual orientation. Many such findings of dissociation have been obtained using the Implicit Association Test (IAT), a procedure described below.[19]

Implicit Stereotypes

A *stereotype* is a mental association between a social group or category and a trait. The association may reflect a statistical reality, but it need not. If the association does exist, it should take the form of members of the group being more likely to display the trait than are members of other groups. A perfect or near-perfect correlation, which might be a *defining trait*, does not fall within the understanding of stereotype. For example, the correlation between race and skin color may be close to perfect but not taken as part of a racial stereotype. Stereotypes are of interest when the correlation between group membership and trait expression is not perfect, but the trait nevertheless distinguishes members of one group from others. For example, perhaps 10–15% of persons over the age of 70 characteristically drive on highways at speeds noticeably below speed limits. Perhaps only 5% of younger persons drive this slowly. If these estimates (which we, in fact, made up) are accurate, they constitute a genuine association between age and driving behavior. This is a stereotype that (as described) applies to only a small minority (10–15%) of elderly people. Nonetheless, it may come to serve as a default assumption — the assumption that *any* elderly person is likely to be a slow driver.

The stereotype that associates male gender with fame-deserving achievement was used in the first experimental demonstration of implicit stereotypes. The experiment

[19] Discussions of factors that promote dissociation of implicit from explicit attitudes can be found in articles

was based on Jacoby et al.'s earlier described false-fame implicit memory effect.[20] Banaji and Greenwald[21] found that this false-fame effect was substantial when the pronounced (i.e., attended-to-but-not-studied) names were male, but was noticeably weaker when the names were female. Banaji and Greenwald described this result as an implicit indicator of the stereotype that associates maleness with fame-deserving achievement. Put more technically, an implicit stereotype of this kind can be defined as "the introspectively unidentified (or inaccurately identified) traces of past experience that mediate attributions of qualities to members of a social category".[22]

Stereotypes can involve associations of favorable or unfavorable traits with a group. Because the favorable–unfavorable distinction is also central to the concept of attitude, it is natural to ask how stereotypes and attitudes differ. For stereotypes the content of the ascribed trait, rather than its evaluative valence, is central. For attitudes, the opposite holds. So, for example, in the implicit fame experiment, it was the *trait* of fame, rather than the *positivity* of fame, that defined the implicit stereotype phenomenon. Steerotypes are often a mixture of favorable and unfavorable traits. As an example, the (not necessarily justified) stereotype of a cheerleader includes both being attractive (a favorable trait) and being unintelligent (an unfavorable trait).

---

by Hofmann et al. (2005) and Nosek (in press).

[20] Jacoby et al. (1989)

[21] Banaji and Greenwald (1995).

[22] Greenwald and Banaji (1995, p. 15)

3. Response Bias and Implicit Bias

The term, "bias," sometimes referred to as "*response bias*," denotes a displacement of response along a judgment continuum. Response bias need not indicate something unwise, inappropriate, or even inaccurate. For example, instructors may vary in their response bias in grading, such that some assign a relatively higher grade to average student performance and others assign a relatively lower grade to the same performance. Instructors who differ in response bias on the grading dimension may nevertheless be equally sensitive to differences among students. Consider an instructor who is biased to grade leniently and assign grades exclusively between A (highest) and C (lowest). This instructor's grades may be perfectly correlated with those of a severe-grading instructor who limits grades to the B to D range. If these two instructors were grading the same person, the more lenient instructor's grades would be exactly one letter grade above those of the more severe instructor. Unless there are established standards that associate specific performances with specific grades, one could not accuse either instructor of being less "accurate" than the other.

A more widely recognized form of bias does affect response accuracy and bears a pejorative connotation. Imagine that a particular instructor differentially assigns grades to two identically performing students when one student is male and the other is female, or when one is White and the other is Black. In this case, judgment fairness and accuracy are both compromised. Attitudes and/or stereotypes are one plausible cause of such *discriminatory* biases. If, among equally qualified job applicants, one favors members of one race over those of another, this plausibly reflects an attitudinal bias —

one may have a more favorable attitude toward one race group than toward the other. If, among equally qualified renters, one assumes that members of one race are more conscientious than those of another, this is plausibly a bias rooted in stereotype. If, among equally qualified candidates for a management position, men are considered preferable to women, it could be due to operation of a stereotype that treats leadership as a trait more frequently found among men than women.[23]

Implicit biases are discriminatory biases based on implicit attitudes or implicit stereotypes. Implicit biases are especially intriguing, and also especially problematic, because they can produce behavior that diverges from one's avowed or endorsed beliefs or principles. The mere existence of implicit bias poses a challenge to legal theory and practice, because discrimination doctrine is premised on the assumption that, barring insanity or mental incompetence, human actors are guided by their avowed (explicit) beliefs, attitudes, and intentions.[24]

Biases can be either favorable or unfavorable. *Ingroup bias* designates favoritism toward groups to which one belongs. There is a widespread intuition that it is often entirely acceptable to be biased in favor of at least some of the groups to which one belongs. In this view, bias is a problem only when it is directed *against* some group. Thus it may be considered acceptable to be biased in favor of one's siblings, children, or other relatives, and similarly to favor one's schoolmates and other friends.

---

[23] Discriminatory biases are plausibly stereotype-based when they oppose the bias that might be expected as an attitude effect. For example, gender biases that discriminate against women are plausibly stereotype-based, given that research has found that attitudes toward women are often more favorable than attitudes toward men (Eagly and Mladinic, 1989).

[24] See generally in this symposium, Linda Hamilton Krieger & Susan T. Fiske, Behavioral Realism in Employment Discrimination Law: Implicit Bias and Disparate Treatment, __ Cal. L. Rev. (2006).

Interestingly, the intuition that biases in favor of one's smaller ingroups (such as family and friends) are acceptable typically does not extend to believing that biases favoring one's larger ingroups (one's race, sex, ethnicity, religion, or age group) are appropriate. Is there a boundary encompassing ingroups toward which favorable biases can be considered acceptable? A non-psychological boundary can be found in the illegality of biases toward certain groups (regardless of one's membership) defined, for example, by race, sex, ethnicity, religion, and age. Psychologically, the small size of some ingroups is no doubt significant because many people feel more obliged to help others when they are one of only a few people who can possibly be helpful,[25] as when family members are the only ones in a position to help.

Perhaps fortunately, the situations in which one wishes to be biased in favor of one's important and smaller ingroups, such as in providing care for one's children, are often ones for which there is no question of possible discrimination against others. Nevertheless, a positive attitude toward any ingroup necessarily implies a *relative* negativity toward a complementary outgroup. In some circumstances, this relative favoring of the ingroup puts members of other groups at a clearly discriminatory disadvantage, as when one allows favoritism toward a family member or friend to influence a hiring, job assignment, rental, or admissions decision in which one plays a role.

[25] This psychological truth was demonstrated very clearly by Darley and Latané (1968) who found that a solitary witness to a simulated epileptic seizure was considerably more likely to intervene than was one of a group of witnesses. This effect of being in a unique position to help is so strong that the presence of multiple bystanders can result in less likelihood of any help being given than when only a single bystander is present.

4. The Implicit Association Test

The recent development of the Implicit Association Test (IAT) has noticeably accelerated research on implicit bias. The IAT's general method can be readily adapted to measure a wide variety of the group–valence and group–trait associations that underlie attitudes and stereotypes. The IAT is an implicit measure because it infers group–valence and group–trait associations from performances that are influenced by those associations, and does so in a fashion that is not discerned by respondents.[26]

The most widely used IAT measure assesses implicit attitudes toward African Americans (AA) relative to European Americans (EA).[27] In this "Race IAT", respondents first practice distinguishing AA from EA faces by responding to faces from one of these two categories with the press of a left-side computer key and to those of the other category with a right-side key. Respondents next practice distinguishing pleasant-meaning from unpleasant-meaning words in similar fashion. The next two tasks, given in randomly determined order, use all four categories. In one of these two tasks, the AA faces and pleasant words get one response (say, left hand) while the EA faces and

[26] The IAT was first reported by Greenwald, McGhee, and Schwartz (1998). The name of the IAT may suggest that it measures something that should be called "implicit associations". However, the "implicit" in the name of the technique designates a type of test, not a type of association. The IAT is therefore a procedure that provides an implicit measure of association strengths. The statement that respondents do not discern the influence of associations on their IAT performance is properly limited to respondents who have not become aware of the way in which the procedure captures association strengths.

[27] These formal race category labels are used in the Race IAT — in place of the widely used labels of Black and White — because the color-name labels carry race-unrelated associative connotations of good and bad that potentially interfere with their use in the measure of race–valence associations.

unpleasant words get the other response. In the remaining task, EA faces share a response with pleasant words and AA faces with unpleasant words.[28]

The implicit attitude measure produced by this IAT is based on relative speeds of responding in the two four-category tasks. This measure allows an inference about attitudes (category–valence associations) because it is easier to give the same response to items from two categories when those two categories are associated than when they are not. For American respondents taking the Race IAT, performances very often take the form of faster performance when EA and pleasant share a response than when AA and pleasant share a response. This frequently observed pattern supports the interpretation that EA–pleasant is a stronger association than AA–pleasant. This result is described as showing implicit attitudinal preference for EA relative to AA.[29]

Research comparing IAT (implicit) measures with parallel standard survey-type self-report (explicit) measures has found systematic variations in the agreement between these two types of measures. There is substantially greater overlap when implicit and explicit attitudes have been shaped by the same experiences, which is likely to be the case for attitudes toward consumer brands, sports teams, and political candidates.[30] When implicit and explicit measures of attitudes or stereotypes disagree — for example,

[28] Various non-essential aspects of the IAT procedure, such as hand assigned to the pleasant category and order of performing the two four-category tasks, are randomized or counterbalanced to avoid their systematically influencing findings.

[29] Because each task involves two associations, the complete description of the inference about association strengths is that the combined strength of EA–pleasant and AA–unpleasant associations is stronger than the combined strength the of AA–pleasant and EA–unpleasant associations. This association-strength interpretation of the IAT has been widely, although not uniformly, accepted. A recent discussion of alternative interpretations can be found in Nosek, Greenwald, & Banaji (in press). One of these alternative interpretations — that the IAT measures cultural beliefs — is considered in Section 5 of this article.

[30] Hofmann et al. (2005), Nosek (in press), Greenwald, Nosek, and Banaji (2003).

when a Race IAT shows preference for EA and a self-report measure shows impartiality — there is said to exist a dissociation between the two.

5. Predictive Validity of the IAT

Researchers have been extending the IAT into increasingly diverse content domains, applying its general method to a wide variety of groups and social categories.[31] Perhaps because of the importance of attitude as a theoretical construct in psychology, more attention, thus far, has been given to investigating implicit attitudes than to investigating implicit stereotypes. In recognition of the importance of understanding relations between IAT measures and behavior, many studies that have used an IAT attitude measure have also included a measure of one or more social behaviors that are theoretically expected to be related to attitude or stereotype measures. Data analyses then determine whether individual differences in implicit attitudes or stereotypes measured by the IAT correlate as expected with individual differences in behavior.

Sixty-one such studies were recently identified and summarized in a meta-analytic review by Poehlman, Uhlmann, Greenwald, and Banaji.[32] The meta-analytic method of their review appraises the value of IAT measures by looking at this body of research in the aggregate, rather than as isolated research findings. To do this, the available correlational measures of relations between IAT and behaviors of interest are averaged within groups of studies that test related hypotheses, as well as over the entire group of

[31] See Nosek (in press).

86 independent findings that were available in the 61 studies. For comparison, parallel analyses were done examining the aggregated correlations involving self-report (explicit) measures, which were also obtained in most of the studies. Both the implicit (IAT) and the parallel explicit measures displayed *predictive validity*, meaning that both types of measures, on average, were significantly correlated in expected fashions with measures of behavior. To be clear, this does not mean that statistically significant correlations were found in all studies, but that averages of the correlational results of the collected similar tests clearly showed the expected positive relationships. Predictive validity was greater (meaning that the average correlation was larger) for self-report (explicit) measures than for IAT measures. However, within the critical group of studies that focused on prejudicial attitudes and stereotypes — in other words, studies of implicit bias — *predictive validity was significantly greater for the IAT measures*.[33]

6. How Pervasive Is Implicit Bias?

IAT measures of implicit attitudes have been available on the internet for self-administered demonstration use since 1998.[34] These web-accessible demonstrations, which provide opportunities to interactively experience the IAT, have accumulated sufficient data to allow conclusions about relative pervasiveness of implicit and explicit

[32] Poehlman, Uhlmann, Greenwald, & Banaji (2005)
[33] Poehlman et al.'s (2005) meta-analysis also found that predictive validity of IAT measures was relatively unaffected by two types of study variations that affected predictive validity of explicit measures. First, predictive validity of explicit measures was reduced when the topic was socially sensitive, such that research subjects might be reluctant to describe their attitudes and beliefs to researchers (e.g., for the domain of race attitudes). Second, predictive validity for explicit measures was likewise reduced when the actions or judgments being observed were "spontaneous", meaning that the actions were ones typically performed with little deliberation or with low levels of conscious control.
[34] Interactive demonstrations of a dozen versions of the IAT are available at https://implicit.harvard.edu.

biases.[35] Table 1 displays results for a dozen data sets, comparing the degree of favoritism toward advantaged (relative to disadvantaged) groups revealed by implicit and explicit measures. Two differences between the implicit and explicit measures are readily apparent in these data. First, the explicit measures generally show much greater evidence for attitudinal neutrality or impartiality. Averaged across the dozen topics, 42% of respondents expressed neutrality on explicit measures, compared to only 18% of respondents on IAT measures being judged to show negligible bias in one or the other direction. Second, the IAT measures consistently showed greater evidence for bias in favor of the relatively advantaged group (averaging 73% across topics) than did the explicit measures (averaging only 38% favoring advantaged groups). Table 1 also shows a bias index computed as the percentage of respondents showing favorability to the advantaged group minus that showing favorability to the disadvantaged group. Whereas this index averaged only 20% for explicit measures, it averaged 64% for IAT measures. The broad generalization justified by the data in Table 1 is that implicit attitudes reveal far more bias favoring advantaged groups than do explicit measures.

The data in Table 1 cannot be interpreted as representing the attitude distribution of some specific population of interest, such as adult residents of the United States. Rather than coming from a random sample of the U.S. population, these data were provided by voluntary visitors to the web site at which demonstration and research versions of IAT measures were available. The drop-in respondents who provided the data in Table 1 therefore constitute *self-selected* samples, as distinct from

[35] These demonstration tests were not set up to conduct research, but were nevertheless obliged to record data to enable feedback of test results to web visitors. The accumulated data provided by the

representative samples that can be obtained by selecting and recruiting respondents randomly from a defined population. Even so, the greater favoritism to advantaged groups found for IAT than explicit measures would certainly also be found with representative samples. The key evidence for this comes from a closer look at the Race IAT data for the wide-ranging demographic subgroups shown in Table 2.

Table 2 shows that, with one notable exception, percentages of respondents who display implicit race bias vary relatively little across groups categorized by age, sex, and educational attainment. The *only* subgroup of respondents who did not show substantial implicit race bias on the Race IAT is African Americans. Approximately equal percentages of African Americans displayed implicit bias in the pro-African American and pro-European American directions. Among African Americans, it was quite noticeable that the IAT results showed considerably greater favoritism to the dominant European American group than did the results from self-report measures, which showed very strong favoritism toward African Americans. The results shown in Table 2 strongly suggest that *any* non-African American subgroup of the United States population will be characterized by high proportions of persons showing statistically noticeable implicit race bias in favor of EA relative to AA.

site's many visitors has proved to be a remarkably rich archive.

Table 1. Distributions of Responding on Self-report (Explicit) and IAT (Implicit) Measures

disadvantaged group	advantaged group	N	% biased toward disadvantaged (dis) and advantaged (adv) groups, and % neutral (neu)							
			Self-report (Explicit)				IAT (Implicit)			
			dis	neu	adv	index	dis	neu	adv	index
IAT Demonstration Web Site Tests										
Afr. American	Eur. American	22074	11.3%	54.6%	34.0%	23%	10.6%	24.1%	65.3%	55%
Old	Young	11528	16.7%	36.8%	46.6%	30%	4.7%	14.3%	81.1%	76%
IAT Research Web Site Tests										
Afr. American	Eur. American	211	11.8%	56.4%	31.9%	20%	12.3%	18.5%	69.2%	57%
Asians	Whites	211	16.4%	56.9%	26.7%	10%	11.3%	25.9%	62.8%	51%
Canadian	American	218	24.1%	39.5%	36.4%	12%	13.3%	21.7%	65.0%	52%
Foreign places	American places	178	20.9%	36.6%	42.4%	22%	9.6%	14.0%	76.4%	67%
Gay people	Straight people	217	14.3%	45.7%	40.0%	26%	8.3%	22.9%	68.8%	60%
Muslims	Jews	144	10.4%	49.3%	40.3%	30%	11.1%	20.7%	68.2%	57%
Old people	Young people	236	27.4%	39.2%	33.5%	6%	5.5%	15.6%	78.9%	73%
Poor	Rich	211	36.7%	37.6%	25.7%	-11%	1.4%	4.3%	94.3%	93%
Fat people	Thin people	239	13.4%	42.4%	44.2%	31%	13.1%	20.8%	66.1%	53%
Japan	USA	263	19.9%	19.9%	60.2%	40%	6.2%	15.2%	78.7%	73%
AVERAGES (12 data sets, unweighted)			19.5%	42.4%	38.1%	20%	9.2%	18.0%	72.8%	64%

Note. Implicit data were obtained with IAT measures (see Section 4) in which pleasant and unpleasant words were classified together with the items representing the groups shown in the table. Explicit data were obtained with self-report measures of endorsed attitudes. The “index” column reports a bias index computed as the percentage favoring the advantaged group minus the percentage favoring the disadvantaged group. The higher the value of this index, the more pervasive is bias. The bias index’s values for IAT measures reveal considerably higher values than for the self-report measures, indicating that implicit bias is far more pervasive than explicit bias. Data from the demonstration web site are previously unpublished. Those from the research web site were reported by Nosek (in press).

Table 2. Percentages Favoring European American (EA) Relative to African American (AA) on Self-report (Explicit) and IAT (Implicit) Measures

Subcategories	N	Self-report (Explicit)				IAT (Implicit)			
		Percent favoring				Percent favoring			
		AA	neither	EA	Index	AA	neither	EA	Index
Education Level									
thru high school grad	3869	9.9%	57.9%	32.2%	22%	9.8%	26.2%	64.0%	54%
at least some college	13028	11.3%	54.1%	34.6%	23%	10.2%	23.2%	66.6%	56%
at least some grad school	3829	12.5%	53.5%	34.0%	21%	12.4%	24.8%	62.9%	50%
Race and Ethnicity									
Black (incl. multiracial)	2277	58.9%	36.2%	4.8%	-54%	34.1%	33.6%	32.4%	-2%
Hispanic (not Black)	1204	15.0%	59.7%	25.3%	10%	10.2%	29.2%	60.5%	50%
Asian & Pacific Islander	1080	9.6%	57.5%	32.9%	23%	7.7%	24.8%	67.5%	60%
White	14805	3.4%	56.0%	40.7%	37%	6.8%	21.7%	71.5%	65%
Age									
under 25	13823	9.7%	55.7%	34.5%	25%	9.4%	23.7%	66.9%	58%
25-44	5403	14.9%	53.9%	31.2%	16%	12.8%	24.4%	62.8%	50%
45 and older	1743	12.3%	47.1%	40.6%	28%	12.6%	25.6%	61.8%	49%
Sex									
Female	13060	12.3%	57.8%	29.8%	17%	11.4%	25.2%	63.4%	52%
Male	7971	9.6%	49.4%	41.0%	31%	9.2%	22.2%	68.6%	59%
Political Ideology									
Conservative	3053	4.8%	44.0%	51.2%	46%	6.5%	19.9%	73.6%	67%
Middle	10612	11.0%	54.0%	35.0%	24%	10.3%	23.8%	65.9%	56%
Liberal	6427	14.8%	59.9%	25.3%	11%	12.9%	26.0%	61.1%	48%

See the note for Table 1. The finding of high levels of the bias index for all demographic subgroups other than Black (i.e., African American) indicates the pervasiveness of pro-European American bias. Even though the bias index was lower in groups of Hispanics and political liberals than for other groups, it was still quite high among those groups.

7**. Why is implicit bias so pervasive?**

This question can be decomposed into three other questions: First, how are implicit attitudes and stereotypes acquired? Second, what mental representations underlie implicit attitudes and stereotypes? Third, do the representations underlying implicit attitudes and stereotypes differ from those underlying explicit attitudes and stereotypes? Answers to these questions could explain both the weak correlations often observed between IAT and explicit measures and the substantially greater bias apparent on implicit than explicit attitudes. It may be several years before thorough research-based answers to these questions are available. These answers will require, in part, research that examines the formation of implicit attitudes and stereotypes in young children. To be used with pre-school children, the IAT needs modifications, the most substantial of which is to replace printed-word stimulus items either with pictures or with words presented in spoken rather than printed form.[36]

In a recent review article, Rudman wrote "The hypothesized causal influences on attitudes include early (even preverbal) experiences, affective experiences, cultural biases, and cognitive consistency principles. Each may influence implicit attitudes more than explicit attitudes, underscoring their conceptual distinction."[37] Rudman's proposal that early experiences and affective experiences may be reflected more in implicit attitudes than in explicit attitudes may explain why implicit attitudes generally reveal more bias, as shown in Tables 1 and 2. As Rudman also noted, influences of cultural factors on the IAT can also explain why people often display implicit attitudes that

[36] Research with IAT procedures that have been adapted for use with pre-school children is now actively under way in the laboratories of M. R. Banaji at Harvard University and A. L. Meltzoff at University of Washington.

appear more concordant with their general cultural milieu than with the experiences of their individual upbringing. One example of this phenomenon is the finding, evident in Table 2, that implicit racial attitudes of African Americans, on average, show much less ingroup-favoring bias than is observed in implicit race attitudes of European Americans.[38]. On average, African Americans' implicit racial attitudes are remarkably close to indicating racial neutrality.[39] And, as was already suggested in considering Table 2, this observation contrasts sharply with findings for African Americans' explicit racial attitudes, which are strongly polarized in the ingroup-favorable (pro-AA) direction.

The finding of approximate racial neutrality of African Americans' implicit attitudes is especially impressive because it is such a clear exception to the general pattern (see Tables 1 and 2) of implicit attitudes revealing more bias than explicit attitudes. Perhaps the appropriate principle needed to accommodate this observation is that, compared to explicit attitudes and explicit stereotypic beliefs, implicit attitudes and stereotypes are more likely to reflect the social–cognitive content of the larger culture. This suggestion leads to an interesting puzzle: If implicit attitude and stereotype measures are indicators of the social–cognitive content of one's broad cultural environment, then Table 1's data indicate that (for as-yet-unclear reasons) explicit measures of attitudes and stereotypes do not reflect the social–cognitive content of the culture of those who provide the measures.

The just-stated proposition that "explicit measures of attitude and stereotype do not reflect the social–cognitive content of the culture of those who provide the measures" is

---

[37] Rudman (2004, p. 79)

[38] See discussion of this and related observations in the discussion of system justification in this symposium's article by Blasi and Jost.

certainly a discouraging assessment of the value of explicit measures. It also provides a perspective on one of the most reasonable and plausible critiques that has been offered in regard to IAT measures. The essence of this critique is that IAT measures should be interpreted, not as indicating beliefs or attitudes that respondents personally endorse but, rather, as indicating modal beliefs or attitudes that respondents understand to be generally endorsed by others — in other words, *cultural beliefs*.[40]

In viewing the IAT as a measure of one's understanding of cultural beliefs that are external to one's self, the critiques just mentioned imply that individual differences in IAT measures are differences in clarity of one's perception of modal cultural beliefs. As such, they might be seen as having no more importance as determinants of interesting forms of social behavior than should differences among people in their clarity of other perceptions, such as the symbols on an eye chart. In contradiction of this expectation, however, the previously mentioned meta-analytic evidence for predictive validity of IAT measures[41] clearly shows that IAT measures do predict social behavior, thereby suggesting an important weakness of the cultural-belief interpretation of IAT measures.[42]

---

[39] see also Nosek, Banaji, and Greenwald (2002, p. 105).

[40] Olson and Fazio (2004) describe these as “extrapersonal associations”; Karpinski and Hilton (2001) label them as “environmental associations”; and Arkes and Tetlock (2004) refer to them “shared cultural knowledge”.

[41] Poehlman et al. (2005); see Section 5, above

[42] Consider the intellectual puzzle created by the collection of views of those who regard the IAT as reflecting cultural beliefs rather than implicit attitudes. Two of these views are that (a) the IAT reflects cultural beliefs, and (b) the IAT measures something different from what is assessed by explicit measures. It follows (logically) that (c) explicit measures do not measure cultural beliefs. Another belief endorsed by many, including those who advocate the cultural-beliefs critique of the IAT, is that (d) explicit measures assess views that respondents avow or endorse. Juxtaposing (c) and (d), one arrives (again, logically) at the seemingly paradoxical conclusion that people’s endorsed beliefs do not correspond to cultural beliefs. This is paradoxical in approximately the same sense that it is paradoxical for the great majority of people to believe that they are above average in intelligence. It is genuinely puzzling to arrive at the conclusion that explicit measures assess something other than cultural beliefs. What do explicit

8. What Is The Difference Between Possessing Implicit Bias And Endorsing Discrimination?

Many who have taken race IAT measures have received their result in the form of the specific statement, "You show automatic preference for European American relative to African American". Those receiving this result frequently ask, "Does this mean that I'm prejudiced?" The answer depends on one's understanding of the difference between what is measured by the IAT and what is measured by explicit (self-report) measures. The answer, "Yes, your IAT result means that you are prejudiced" would follow from believing that the IAT provides a measure of racial attitude that is at least as appropriate as a measure of prejudice as is the measure provided by self-report. The answer "No, your IAT result has relatively little bearing on a judgment of whether or not you are prejudiced" follows from considering the likelihood that IAT and self-report measures assess distinct mental constructs.

Because psychological interpretations of research data are remarkably flexible, it is possible to accommodate the numerous findings of dissociation of IAT from explicit measures[43]. with widely diverging theoretical interpretations. At one extreme, one may assume that IAT and explicit measures correspond to independent sources of knowledge that are housed in separate brain regions. At the other extreme, one can assume that the two types of measures are based on exactly the same stored knowledge. With the latter assumption, discrepancies between results obtained with the

---

measures assess if not the beliefs that are dominant in one's culture? The implausibility of these logical consequences of the cultural-belief critique constitutes a problem not yet addressed by the critique's proponents.

[43] See definitions of dissociation in Sections 1 and 4, above

two types of measures are assumed to be due to influences on behavior that operate very differently in the research settings in which the two types measures are obtained.[44]

Because of the range of available interpretations, the question of whether or not the IAT measures prejudice has no single answer. From the perspective of a *dual-representation* view, which treats IAT and explicit measures as corresponding to mentally distinct entities, implicit attitudes may not measure prejudice. In this view, prejudice is associated primarily with the portion of the psyche that is responsible for expressions of attitudes observed on explicit (self-report) measures. In sharp contrast, advocates of a *single-representation* view are prepared to describe implicit race preference as indeed corresponding to what is ordinarily meant by "prejudice".

At present, the dual-process view appears to be more widely held among social cognition researchers, as suggested by their frequent descriptions of the weak relation between implicit and explicit attitude measures as indicating "dissociation". In this dominant dual-process view, those who show stronger association (for example) of European American than African American with pleasant are not prejudiced in the ordinary meaning of having an avowed racial preference. Rather, this frequent result of the race IAT is often described as indicating "implicit prejudice" or "automatic preference". These uses of the qualifiers "implicit" and "automatic" are intended to differentiate what the IAT measures from the lay-language meaning of "prejudice", which is assumed to be something more properly assessed via self-report measures.

[44] Fazio (1990)

9. Do Implicit Biases Produce Discriminatory Behavior?

As noted earlier, evidence indicating that implicit attitudes produce discriminatory behavior is already substantial.[45]. Such evidence will almost certainly continue to accumulate. The dominant interpretation of this evidence is that implicit attitudinal biases are especially important in influencing non-deliberate or spontaneous discriminatory behaviors.

A study by McConnell and Leibold,[46] one of the first to relate an IAT race attitude measure to discriminatory behavior, provides a good illustration. In this study, the behavior of White undergraduate students was videotaped while they were being interviewed both by a White and a Black experimenter.[47]. These subjects also completed a race attitude IAT measure. Subjects whose race IAT scores indicated strong implicit preference for White relative to Black hesitated less and made fewer speech errors when speaking to the White experimenter than the Black experimenter. They also spoke more to and smiled more at the White experimenter than the Black experimenter. These behaviors suggest higher levels of comfort interacting with the White experimenters.[48]

The importance of results such as McConnell and Leibold's becomes especially apparent in light of the findings of research that was conducted by Word, Zanna, and

[45] Poehlman et al. (2005). See previous mentions of this in Sections 5 and 7.
[46] McConnell and Leibold, (2001)
[47] Subjects did not know in advance about the videotaping, but received a subsequent description and explanation, at which time their signed consent to use the videotape as a source of research data was obtained
[48] Other published studies have likewise found correlations of IAT-measured race attitude with indicators of subtle or spontaneous discriminatory actions. These include Ashburn-Nardo, Knowles, and Monteith (2003); Hugenberg & Bodenhausen (2003, 2004); and Richeson et al. (2003). Several similar unpublished results involving race IAT measures were mentioned in the Poehlman et al. (2005) meta-

Cooper[49] well before development of the IAT. In the first of two studies, Word et al. showed that, in interviewing both a Black and White job applicant, White students showed greater indications of nonverbal discomfort and spent less time speaking with the Black applicant — precisely the same indicators that McConnell and Leibold found were predicted by individual differences on their IAT measure of implicit race attitude. In Word et al.'s second study, White interviewers were carefully trained to control these same subtle aspects of their behavior in interaction with naive, White job applicants. Compared to those whose interviewer behavior resembled the relatively comfortable behavior of the White–White interactions in the first study, Interviewees who encountered the less-comfortable interviewer performed less well in the interview and were more uncomfortable and distant in their interaction style. They also judged their interviewer to be less friendly. The combination of the McConnell–Leibold and Word–Zanna–Cooper findings reveals the potential for implicit bias to affect interaction quality in ways that can disadvantage Blacks.

Another noteworthy result showing prediction by the IAT is the finding that the Race IAT, administered to White American subjects. predicts activation of the amygdala — an indication of fear or other negative emotional arousal — in response to photographic images of unfamiliar African American faces.[50] A related finding was the report by Richeson et al.[51] that IAT measures correlated with evidence of self-regulatory or executive control activity on exposure to African American faces.

---

analytic review, along with similar results from studies of implicit biases toward other stigmatized groups, such as Germans' implicit attitudes toward Turks.

[49] Word, Zanna, and Cooper (1974)

[50] Phelps, O'Connor, Cunningham, Funayama, Gatenby, Gore, and Banaji (2000).

[51] Richeson, Baird, Gordon, Heatherton, Wyland, Trawalter, and Shelton (2003).

10. <u>What can be done to attenuate the influence of implicit biases on behavior?</u>

In their 1995 review of then-available research evidence, Greenwald and Banaji suggested that attentional focus would attenuate automatic influences on social judgment, if those influences were relatively weak.[52] Applying this principle, and assuming that implicit biases constitute "weak automatic influences", one might expect to eliminate their influences on interpersonal behavior by getting people to think more about, or to attend more closely to, their objectives in an interracial interaction. However, Poehlman et al.'s[53] review of the relevant predictive validity evidence for IAT measures suggests some limitation of this conclusion. Their review indicated that prediction of behavior by IAT measures was *not* reduced when the behavior under investigation was more deliberative. They also found that predictive validity of explicit measures was greater for more deliberative behavior (i.e., for behavior presumably done with greater self-conscious attention). Applying these findings to a hypothetical situation in which racially different applicants are being evaluated for jobs, for admissions, for loans, for medical treatments, etc., an interviewer's devoting more deliberate effort to the task may increase the interviewer's use of explicit evaluative criteria — and might thereby produce better decisions — but may not eliminate effects of implicit biases. Thus, the meta-analysis's conclusions suggest caution in assuming that implicit bias should be reduced by increased deliberative effort on a decision. Because no studies have yet provided direct experimental tests of this hypothesis, the

[52] Greenwald and Banaji (1995), p. 17.
[53] Poehlman et al. (2005)

question of how to attenuate the impact of implicit biases on subtle-but-important aspects of interpersonal interaction still awaits answer.

11. How Can Implicit Biases Be Altered?

In the first few years after development of the IAT, many researchers working with the technique were impressed that, when they repeatedly administered the same IAT to themselves, their measures of implicit bias remained remarkably similar over time. This was in part a welcome observation, because it indicated that IAT measures might be administered multiple times to the same person without losing their validity as research measures.[54] The consistency of IAT measures over time also suggested that the implicit attitudes and stereotypes measured by the IAT were stable entities that were not easily modified. Subsequent research has shown that conclusion to be premature, as can be illustrated by one of the first experiments that sought to influence IAT performance.

Taking as their cue the assumption that the race–valence associations that are measured by the IAT may be shaped by media exposures, Dasgupta and Greenwald[55] had White undergraduate students complete a preliminary task in which they practiced identifying a series of photographs of well known and admirable African Americans (scientists, artists, political leaders) mixed with photographs of somewhat less well known but thoroughly disreputable European Americans (terrorists and serial murderers). A subsequent race IAT measure revealed that this photograph-identification practice reduced the level of automatic preference for European American

[54] This is a desirable property, which can be noted by its contrast with (for example) the need to have multiple different forms of intelligence tests, for which scores will obviously improve if exactly the same test is administered repeatedly.

(relative to African American). Further, this reduction in implicit bias persisted over a 24-hour delay.[56]

Numerous other such experiments were summarized in a review by Blair[57] who concluded that implicit biases are malleable. That is, Blair found that IAT measures and other measures of implicit bias[58] are modifiable by a variety of procedures that reduce the typical bias. For example, implicit gender stereotypes of feminine weakness were reduced by imagining counter-stereotypic female exemplars,[59] and implicit anti-Black race attitudes were reduced by having the research procedure administered by an African American experimenter.[60] In studies using IAT measures of implicit race attitudes, these effects are most often modest. They typically reduce, but generally do not come close to eliminating, implicit biases.

The question naturally raised by the sizeable collection of findings suggesting that implicit biases might be malleable is, "How durable are these changes?" Although the research needed to answer this question with confidence has not yet been done, most IAT researchers are skeptical that targeted interventions, as opposed to profound social and cultural change, will have enduring effects on levels of implicit bias. Some researchers, for example, suggest that interventions like those used in the malleability studies produce short term changes in implicit bias levels because they temporarily

---

[55] Dasgupta & Greenwald (2001)
[56] The opposite type of preliminary exposure, consisting of photographs of admirable European Americans and disreputable African Americans, had no noticeable impact on race IAT scores. This observation suggested that the ordinary media environment encountered by these undergraduate research subjects might have been functioning as the equivalent of these biased (anti-Black) exposures — cf. Kang (2005).
[57] Blair,(2002),
[58] e.g., Blair & Banaji (1996); Kawakami, Dovidio, Moll, Hermsen, & Russin (2000)
[59] Blair, Ma, and Lenton (2001).
[60] Lowery, Hardin, and Sinclair (2001).

activate a *sub-type* of a larger category, such that this sub-type temporarily replaces the larger category. For example, in the Dasgupta and Greenwald experiment[61] the preliminary exposure to admirable Blacks may, for at least for some of the subjects, cause the relatively attractive sub-type of *African American celebrities* to be activated and temporarily function as a mental replacement for the larger (and presumably negatively valenced) African American category. If this interpretation is correct it seems unlikely that even repeated interventions will produce cumulative effects, if these repetitions occur within a larger societal environment that reinforces preexisting racial attitudes and stereotypes.

This skeptical appraisal does not imply that long-term changes in implicit biases are impossible. For example, research has shown that when a person forms a new personal connection with a member of a previously de-valued outgroup, implicit attitudes toward that group may change dramatically and rapidly.[62] For example, when a son or daughter marries a member of a racial or ethnic minority or when a close friend is paralyzed in an accident and becomes wheelchair-bound, a favorable implicit attitude may rapidly replace a pre-existing negative implicit bias.[63]

---

[61] Dasgupta & Greenwald (2001)
[62] [Greenwald et al. (2002)
[63] Olsson, Ebert, Banaji, and Phelps, (2005) recently reported that an implicit indicator of expected anti-outgroup racial bias was absent for college student subjects who had interracial dating experience. A remarkable model for this type of influence was given by George Orwell in a scene from 1984 (Part 2, Chapter 9). After 20 minutes of haranguing a crowd of Oceanians with vilification of the Eurasian enemy, the orator receives a piece of paper and "without pausing in his speech" continues his tirade against the (new) enemy, Eastasia. Orwell's text continues: "Without words said, a wave of understanding rippled through the crowd. Oceania was at war with Eastasia! . . . The banners and posters with which the square was decorated were all wrong! . . . There was a riotous interlude while posters were ripped from the walls, banners torn to shreds and trampled underfoot. . . . But within two or three minutes it was all over. . . . The Hate continued exactly as before, except that the target had changed."

Another way in which malleability of implicit bias has been tested with IAT measures has been to ask research subjects to try to respond to the IAT so as to produce a specific result — for example, asking subjects who ordinarily show implicit preference for European American to attempt to produce an IAT result that would be scored as indicating the reverse automatic preference (i.e., to produce an IAT result showing preference for African American). Relatively few subjects succeed at this faking assignment, partly because few can spontaneously come up with a faking strategy.[64] However, somewhat successful faking of the IAT has been achieved by instructing subjects to deliberately respond slowly on the IAT task for which they can ordinarily respond rapidly.[65]

12. Is Implicit Bias A Probable Cause Of Disparate Outcomes?

> "When you have eliminated all which is impossible, then whatever remains, however improbable, must be the truth."
>
> Sherlock Holmes, in A. Conan Doyle's "The Adventure of The Blanched Soldier" (1926)

The argument that implicit bias is a probable cause of race discrimination sometimes requires inference by a process of elimination. This is a reasoning device endorsed not only by Sherlock Holmes, but also by the United States Supreme Court. Specifically, in *Furnco v. Waters*,[66] a 1978 employment discrimination case, the Court wrote:

---

[64] Banse, Seise, & Zerbes (2001); Egloff & Schmukle (2002); Steffens (2004)

[65] cf. Kim (2003). Such faking may prove detectable because it is possible to detect that a subject is responding unusually slowly in a task. By comparison, it is generally quite easy for subjects to fake attitudes and beliefs on self-report measures. This typically requires no more than modifying the position on which a pencil mark is placed in giving a response to a survey questionnaire.

[66] 438 U.S. 567 (1978).

[We know from our experience that more often than not people do not act in a totally arbitrary manner, without any underlying reasons . . .Thus, when all legitimate reasons for [a negative outcome] have been eliminated as possible reasons for the employer's actions, it is more likely than not the employer, whom we generally assume acts only with *some* reason, based his decision on an impermissible consideration such as race.[67]

Whether adjudicating an individual allegation of discrimination, or attempting to understand broad patterns of disadvantage in society more generally, if one finds evidence of disparate impact — for example, in the form of systematically disadvantageous outcomes to African Americans in health care, education, employment, housing, or criminal justice — one may begin the process of identifying and eliminating plausible causes. Conceivable explanations that cannot be eliminated remain worth considering.

For sake of argument, let us assume that in attempting to understand whether racial bias has played a role in probation recommendations in a particular criminal court system, all conceivable non-race-related ("racially neutral") explanations have been eliminated through sound research evidence. Let us also assume that none of the relevant decision makers have reported consciously holding negative racial attitudes or stereotypes. Finally, let us assume that no test of implicit bias has been administered to these decision makers. With this set of assumptions, is it reasonable to infer that the observed racial disparity is being caused by implicit bias? Not only is it reasonable, it

[67] Id. at 577.

should be regarded as the most probable cause. This conclusion is justified by three considerations.

The first consideration is the observed pervasiveness of implicit bias, as was demonstrated very clearly by the data summarized in Tables 1 and 2.

The second consideration comes from the available evidence that (a) implicit biases are predictive of discriminatory behavior and (b) implicit bias measures do a significantly better job than explicit bias measures in predicting behavioral indicators of prejudice or discrimination. This evidence was previously described in Sections 5 and 9.

The third consideration is provided by findings that implicit bias plays a causal role in discrimination. The most important piece of this evidence at present is the finding (described in Section 9) that subtle discriminatory behaviors, of the types known to be predicted by IAT measures of implicit race bias, play a significant role in determining the outcomes of job interviews.[68] The absence of another type of evidence also supports this causal interpretation. Specifically, if — in the joint absence of racially neutral causes and explicit bias — racially disparate impact is shown to occur even when implicit bias is shown to be absent, this would provide evidence against a causal role of implicit bias in disparate impact. No such evidence now exists.

In summary, a substantial and actively accumulating body of research evidence establishes that implicit race bias is pervasive and that implicit race bias is associated with discrimination against African Americans. Consequently, it is plausible that, when racially neutral causes and explicit bias can be rejected as plausible causal

[68] see especially McConnell and Leibold (2001); Word, Zanna, and Cooper, (1974).

explanations for racially disparate outcomes, one can reasonably infer that implicit race bias played a determinative causal role. More direct confirmations of the causal role of implicit bias can be expected in the next few years as studies in relevant domains in which racially disparate impact is a known phenomenon (e.g., health care, education, employment, housing, and criminal justice) increasingly include measures of implicit bias.

References

Arkes, H. R., & Tetlock, P. E. (2004). Attributions of implicit prejudice, or "Would Jesse Jackson 'fail' the Implicit Association Test?" *Psychological Inquiry*. *15*, 257–278.

Ashburn-Nardo, L., Knowles, M.L., & Monteith, M.J. (2003). Black Americans' implicit racial associations and their implications for intergroup judgment. *Social Cognition*, *21*, 61-87.

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

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., Gollwitzer, P. M., Lee-Chai, A. Y., Barndollar, K., & Troetschel, R. (2001). The automated will: Nonconscious activation and pursuit of behavioral goals. *Journal of Personality and Social Psychology*, *81*, 1014–1027.

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

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

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

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences* (Rev. ed.). New York: Academic Press.

Darley, J. M., & Latané, B. (1968). Bystander intervention in emergencies: Diffusion of responsibility. *Journal of Personality and Social Psychology*, *8*, 377–383.

Dasgupta, N., & Greenwald, A. G. (2001). Exposure to admired group members reduces automatic intergroup bias. *Journal of Personality and Social Psychology*, *81*, 800-814.

Eagly, A. H., & Mladinic, A. (1989). Gender stereotypes and attitudes toward women and men. *Personality and Social Psychology Bulletin, 15*, 543-558.

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.

Erdelyi, M. H. (1985). *Psychoanalysis: Freud's cognitive psychology*. New York: Freeman.

Erdelyi, M. H., & Goldberg, B. (1979). Let's not sweep repression under the rug: Toward a cognitive psychology of repression. In J. F. Kihlstrom & F. J. Evans (Eds.), *Functional disorders of memory* (pp. 355-402). Hillsdale, NJ: Erlbaum.

Fazio, R. H. (1990b). Multiple processes by which attitudes guide behavior: The MODE model as an integrative framework. In M. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 23, pp. 75–110). San Diego: Academic Press.

Festinger, L. (1957). *A theory of cognitive dissonance.* Palo Alto, CA: Stanford University Press.

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

Greenwald, A. G., Banaji, M. R., Rudman, L. A., Farnham, S. D., Nosek, B. A., & Mellott, D. S. (2002). A unified theory of implicit attitudes, stereotypes, self-esteem, and self-concept. *Psychological Review*, *109*, 3–25.

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*, 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*, 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*, 197-216.

Heider, F. (1958). *The psychology of interpersonal relations*. New York: Wiley.

Hofmann, W., Gawronski, B., Gschwendner, T., Le, H., & Schmitt, M. (2005). A meta-analysis on the correlations between the Implicit Association Test and explicit self-report measures. *Personality and Social Psychology Bulletin*, *31*, 1369–1385.

Hugenberg, K., & Bodenhausen, G. V. (2003). Facing prejudice: Implicit prejudice and the perception of facial threat. *Psychological Science*, *14*, 640-643.

Hugenberg, K., & Bodenhausen, G. V. (2004). Ambiguity in social categorization: The role of prejudice and facial affect in race categorization. *Psychological Science*, *15*, 342-345.

Jacoby, L. L., Allan, L. G., Collins, J. C., & Larwill, L. K. (1988). Memory influences subjective experience: Noise judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*, 240–247

Jacoby, L. L., & Dallas, M. (1981). On the relationship between autobiographical memory and perceptual learning. *Journal of Experimental Psychology: General***,** *110*, 306–340.

Jacoby, L. L., Kelley, C., Brown, J., Jasechko, J. (1989). Becoming famous overnight: Limits on the ability to avoid unconscious influences of the past. *Journal of Personality and Social Psychology*, *56*, 326–338.

Kang, J. (2005?). Trojan Horses of race. *Harvard Law Review* (need vol. # and pages)

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

Kawakami, K., Dovidio, J. F., Moll, J., Hermsen, S., & Russin, A. (2000). Just say no (to stereotyping): Effects of training in the negation of stereotype associations on stereotype activation. *Journal of Personality and Social Psychology, 78*, 871–888.

Kihlstrom, J. F. (1990). The psychological unconscious. In L. A. Pervin (Ed.), *Handbook of personality: Theory and research* (pp. 445-464). New York: Guilford.

Kihlstrom, J. F., Barnhardt, T. M., Tataryn, D. J. (1992) Implicit perception. In Bornstein, R F., & Pittman, T. S. (Eds.) *Perception without awareness: Cognitive, clinical, and social perspectives*. (Pp. 17-54). New York: Guilford Press.

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

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.

McConnell, A. R., & Leibold, J. M. (2001). Relations among the Implicit Association Test, discriminatory behavior, and explicit measures of racial attitudes. *Journal of Experimental Social Psychology*, *37*, 435–442

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

Nosek, B. A. (in press). Moderators of the relationship between Implicit and explicit evaluation. *Journal of Experimental Psychology: General*.

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

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.

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

Olsson, A., Ebert, J. P., Banaji, M. R., & Phelps, E. A. (2005). The role of social groups in the persistence of learned fear. *Science*, *309*, 785–787.

Orne, M. T. (1962). On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *American Psychologist*, *17*, 776-783.

Orwell, G. *1984.* New York: Harcourt, Brace, 1949.

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

Rosenberg, M. J. (1969). The conditions and consequences of evaluation apprehension. In R. Rosenthal & R. L. Rosnow (Eds.), *Artifact in behavioral research* (pp. 279–349). New York: Academic Press.

Rudman, L. A. (2004). Sources of implicit attitudes. *Current Directions in Psychological Science*, *13*, 79–82.

Rudman, L. A., Greenwald, A. G., & McGhee, D. E. (2001). Implicit self-concept and evaluative implicit gender stereotypes: Self and ingroup share desirable traits. *Personality and Social Psychology Bulletin, 27,* 1164–1178.

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.

Shevrin, H., & Dickman, S. (1980). The psychological unconscious: A necessary assumption for all psychological theory? *American Psychologist*, **34**, 421-434.

Steffens, M. C. (2004). Is the Implicit Association Test Immune to Faking? *Experimental Psychology*, *51*, 2004, 165–179.

Weber, S. J., & Cook, T. D. (1972). Subject effects in laboratory research: An examination of subject roles, demand characteristics, and valid inference. *Psychological Bulletin*, *77*, 273–295.

Wegner, D. M. (2002). *The illusion of conscious will*. Cambridge, MA: MIT Press

Word, C. O., Zanna, M. P., & Cooper, J. (1974). The nonverbal mediation of self-fulfilling prophecies in interracial interaction. *Journal of Experimental Social Psychology*, *10*, 109–120.