Antidiscrimination Law and the Perils of Mindreading

GREGORY MITCHELL*
PHILIP E. TETLOCK**

Recent legal scholarship challenges the default psychological assumption in antidiscrimination law that discrimination is a function of psychological processes under the conscious control of the discriminator, and replaces it with the assumption that discrimination is the result of unconscious, or implicit, psychological processes that operate automatically, beyond conscious control. This challenge is, however, only as persuasive as the research on which it is predicated, and we document that this research fails to satisfy key scientific tests of validity. We conclude that implicit prejudice research should be accepted as neither legislative authority nor litigation evidence until there is more: (1) rigorous investigation of the error rates of the new implicit measures of prejudice (and of how investigators balance Type I errors of false accusations against Type II errors of failing to identify prejudice); (2) thorough analysis of how well implicit measures of prejudice predict discriminatory behavior under realistic workplace conditions; and (3) open debate about the societal consequences of setting thresholds of proof for calling people prejudiced so low that the vast majority of the population qualifies as prejudiced.

I. INTRODUCTION

The shape of the next generation of antidiscrimination law hinges on how legislators and judges respond to the argument that prejudice in America has mutated into new insidious forms—that prejudice, once overt, is now largely covert, indeed, so covert that possessors of the new prejudice are themselves unaware both of the contents of their own minds and of how

---

* Professor of Law & E. James Kelly, Jr.—Class of 1965 Research Professor, University of Virginia School of Law, 580 Massie Road, Charlottesville, VA 22903-1738, greg_mitchell@virginia.edu. We appreciate the helpful comments of Hal Arkes, Hart Blanton, David Copus, Paul Hanges, Brian Kalt, Barbara Mellers, Devah Pager, Gowri Ramachandran, Richard Redding, William von Hippel, Amy Wax, and Jonathan Ziegert on an earlier draft.

** Lorraine Tyson Mitchell Chair of Organizational Behavior, Haas School of Business, University of California, Berkeley, 94720-1900, with formal affiliations with the psychology and political science departments; tetlock@haas.berkeley.edu.

these contents bias their judgments of protected-category groups.[1] Those advancing this argument warn us that survey evidence of declines in prejudice are misleading: prejudice and stereotyping are as robust as ever but now operate more surreptitiously, via unconscious, or implicit, associations among mental categories (White + good/Black + bad, old + feeble/young + healthy) that lead to subtle and not-so-subtle acts of discrimination.[2] Were these insidious associations limited to a small percentage of the population, then wholesale changes to the psychological assumptions of antidiscrimination law would be unwarranted. But if recent claims are to be believed, unconscious processes of discrimination operate pervasively: under this emerging view, most, if not all, of us are implicit bigots most, if not all, of the time.[3]

---

[1] *See* LU-IN WANG, DISCRIMINATION BY DEFAULT 135, 135 (2006) ("[L]egal prohibitions against discrimination are inadequate to redress the largest share of modern discrimination, particularly under the dominant model of intentional discrimination."); Samuel R. Bagenstos, *The Structural Turn and the Limits of Antidiscrimination Law*, 94 CAL. L. REV. 1, 3 ("Unconscious bias, interacting with today's increasingly 'boundaryless workplace,' generates inequalities that our current antidiscrimination law is not well-equipped to solve." (footnote omitted)); Martha Chamallas, *Deepening the Legal Understanding of Bias: On Devaluation and Biased Prototypes*, 74 S. CAL. L. REV. 747, 753 (2001) ("[A]ntidiscrimination law is inadequate because it targets mainly intentional discrimination, missing the more prevalent contemporary forms of bias that are often nondeliberate or unconscious." (footnote omitted)); Susan Sturm, *Second Generation Employment Discrimination: A Structural Approach*, 101 COLUM. L. REV. 458, 460 (2001) ("Cognitive bias, structures of decisionmaking, and patterns of interaction have replaced deliberate racism and sexism as the frontier of much continued inequality." (footnote omitted)); Audrey J. Lee, Comment, *Unconscious Bias Theory in Employment Discrimination*, 40 HARV. C.R.-C.L. L. REV. 481, 482 (2005) ("The nature of discrimination today is dramatically different from the pernicious, overt discrimination that existed prior to the passage of the Civil Rights Act of 1964." (footnote omitted)).

[2] *See, e.g.*, Margo J. Monteith et al., *Taking a Look Underground: Detecting, Interpreting, and Reacting to Implicit Racial Biases*, 19 SOC. COGNITION 395, 396 (2001) ("[W]ith little intent or conscious awareness, negative racial associations that are consciously disavowed can be activated and used as a basis for responding to members of stereotyped groups."); Darren Seiji Teshima, *A "Hardy Handshake Sort of Guy": The Model Minority and Implicit Bias about Asian Americans in* Chin v. Runnels, 11 ASIAN PAC. AM. L.J. 122, 139–40 (2006) ("The notion that racism no longer infects society as a whole, but is rather the intentional act of an anomalous bigot is 'the Big Lie.' . . . Social cognition research has demonstrated that stereotypes affect all of us, even without our awareness of this cognitive process." (footnote omitted)).

[3] Consider, for instance, recent statements to this effect by Jerry Kang & Mahzarin Banaji, who are, respectively, the leading legal and psychological proponents of the implicit prejudice view:

> [B]y a conservative estimate, around ninety percent of Americans (and others in the western world), mentally associate negative concepts with the social group

An explosion of research into the unconscious processes of social cognition provides the empirical foundation for this "implicit prejudice" view.[4] Using new methods that purportedly tap into the subconscious, most prominent among them the Implicit Association Test ("IAT"),[5] some psychologists now claim that the vast majority of the population—including many victims of discrimination—implicitly associate disadvantaged groups

---

> "elderly"; only about ten percent show the opposite effect associating elderly with positive concepts. Seventy-five percent of Whites (and fifty percent of Blacks) show anti-Black bias, and seventy-five percent of men and women do not associate female with career as easily as they associate female to family. . . . These data, as well as the findings in dozens of experiments that meet the criteria of replicability and peer-review, demonstrate that we are not color or gender blind, and perhaps that we cannot be.

Jerry Kang & Mahzarin R. Banaji, *Fair Measures: A Behavioral Realist Revision of "Affirmative Action"*, 94 CAL. L. REV. 1063, 1072 (2006). Kang & Banaji contend further that the new tools of this implicit prejudice research "can measure threats to fair treatment—*threats that lie in every mind*." *Id.* at 1066 (emphasis added).

[4] *See* Patricia G. Devine, *Implicit Prejudice and Stereotyping: How Automatic Are They? Introduction to the Special Section*, 81 J. PERSONALITY & SOC. PSYCHOL. 757, 757 (2001) ("In recent years, there has been a veritable explosion of work on the nature and assessment of implicit components of prejudice and stereotyping. Over the last decade or so, a great many studies have revealed that prejudice and stereotypes can operate without the conscious intent or awareness of social perceivers."). Unless we specify otherwise, when we refer to the "implicit prejudice" line of work, we mean to refer generally to research into unconscious, or implicit, biases toward different social groups, either in the form of implicit prejudicial evaluative attitudes toward these groups or implicit stereotypic beliefs about these groups.

[5] The IAT is one of several devices in social psychology designed to capture attitudes without asking people what views they consciously endorse. For a review of these implicit measures, see Russell H. Fazio & Michael A. Olson, *Implicit Measures in Social Cognition Research: Their Meaning and Use*, 54 ANN. REV. PSYCHOL. 297 (2003); *see also* Anthony G. Greenwald et al., *Measuring Individual Differences in Implicit Cognition: The Implicit Association Test*, 74 J. PERSONALITY & SOC. PSYCHOL. 1464, 1473–74 (1998) (describing the original race IAT). Essentially, the IAT measures millisecond differences in reaction times to pairings of concepts that vary in their putative stereotypic or prejudicial connotations. For instance, if a subject responds more quickly to the pairing of photographs of African-American faces with negative character trait words than to the pairing of photographs of European-American faces with the same negative character trait words, then the subject is said to exhibit an implicit negative stereotype toward African-Americans. Or if a subject responds more quickly to the pairing of "White-sounding" names with the term "pleasant" than to the pairing of "Black-sounding" names with the term "pleasant," then the subject is said to exhibit an implicit negative attitude (or prejudice) toward African-Americans. For a detailed discussion of the IAT procedure and IAT research, see *infra* Part II.B; for a detailed critique of this research, see *infra* Part III.

with negative attributes and advantaged groups with positive attributes.[6] Because these implicit associations are said to arise from basic psychological processes common to all, this new research ostensibly reveals a deep reservoir of prejudice beneath the surface of even seemingly civil intergroup relations.[7] Even more troubling, these implicit associations are said to manifest themselves in acts of discrimination,[8] and good intentions can do little to stop this automatic and unconscious discriminatory process.[9]

Antidiscrimination law scholars have seized on this new research to argue for changes in the legal landscape.[10] Jerry Kang and Mahzarin Banaji,

---

[6] *See* Brian A. Nosek et al., *Harvesting Implicit Group Attitudes and Beliefs From a Demonstration Web Site*, 6 GROUP DYNAMICS: THEORY, RESEARCH, AND PRACTICE 101, 112 (2002) ("From young to old, male to female, Black to White, and conservative to liberal, implicit biases are not held by a select few but are readily observed among all social groups."); *see also* John T. Jost et al., *Non-Conscious Forms of System Justification: Implicit and Behavioral Preferences for Higher Status Groups*, 38 J. EXPERIMENTAL SOC. PSYCHOL. 586, 598 (2002) ("[I]t appears that members of disadvantaged groups internalize negative stereotypes and evaluations of their own group, to at least some degree.").

[7] *See, e.g.*, Max H. Bazerman & Mahzarin R. Banaji, *The Social Psychology of Ordinary Ethical Failures*, 17 SOC. JUST. RES. 111, 111 (2004) ("These ordinary unethical behaviors are conceived to be ordinary because they are assumed to be rooted in the basic mechanics of the mind's abilities and constraints. They are also ordinary in that such unethical behaviors are not characteristic of a special group of unethical people . . . but rather of all of us." (citation omitted)).

[8] *See, e.g.*, Anthony G. Greenwald & Mahzarin R. Banaji, *Implicit Social Cognition: Attitudes, Self-Esteem, and Stereotypes*, 102 PSYCHOL. REV. 4, 7 (1995) ("[A]ttitudes of which the actor is *not* conscious at the moment of action (implicit attitudes) are also strongly predictive of behavior."); Monteith et al., *supra* note 2, at 396 ("[W]ith little intent or conscious awareness, negative racial associations that are consciously disavowed can be activated and used as a basis for responding to members of stereotyped groups.").

[9] *See* Mahzarin R. Banaji et al., *Implicit Stereotyping in Person Judgment*, 65 J. PERSONALITY & SOC. PSYCHOL. 272, 280 (1993) ("Implicit stereotyping critically compromises the efficacy of 'good intention' in avoiding stereotyping and points to the importance of efforts to change the material conditions within which (psychological) stereotyping processes emerge and thrive."); Devine, *supra* note 4, at 757 ("Even those who consciously renounce prejudice have been shown to have implicit or automatic biases that conflict with their nonprejudiced values that may disadvantage the targets of these biases.").

[10] The California Law Review will soon publish a symposium issue devoted to "behavioral realism," with legal implications of implicit prejudice research being a primary topic of discussion, see Symposium, 94 CAL. L. REV. (forthcoming 2006). Further, despite the youth of the Implicit Association Test, at least forty-four law journal articles discuss IAT research in a legal context (as found through a search conducted on Feb. 22, 2006, for "implicit association test" in Westlaw's Journals and Law Reviews

for instance, call for greater use of affirmative action programs and other "fair measures" to counter and change implicit biases against women and minorities.[11] Antony Page calls for new procedures to combat the implicit racial and gender biases likely to influence the exercise of peremptory challenges or, alternatively, outright elimination of the peremptory challenge.[12] And in perhaps the boldest legal application of implicit prejudice research to date, Saujani proposes that the IAT be used to read the minds of legislators for evidence of unconscious discriminatory intent in their enactments.[13]

These specific examples are but a small part of an ambitious project to

---

(JLR) database). We discuss legal scholarship on the IAT in more detail below. *See infra* Part II.C.

[11] *See* Kang & Banaji, *supra* note 3, at 1080 ("[W]e need a new model of discrimination for implicit bias—one based on a more accurate model of human cognition and emotion, especially its constraints. This new model must promote proactive structural interventions that minimize harm without relying solely on potential individual litigation." (footnote omitted)); *id.* at 1116 ("Fair measures that are race- or gender-conscious will become presumptively unnecessary when the nation's implicit bias against those social categories goes to zero or its negligible behavioral equivalent." (footnotes omitted)). The specific "fair measures" Kang and Banaji propose include legal use of tie-breaking in favor of women and minorities in hiring, the hiring of counter-stereotypical women and minorities to serve as "debiasing agents" in organizations, raising self-awareness about implicit biases by having employees take the IAT and other psychological tests of implicit bias, and cloaking social category information in employment decision-making processes. *See id.* at 1101–15.

[12] *See* Antony Page, Batson's *Blind-spot: Unconscious Stereotyping and the Peremptory Challenge*, 85 B.U. L. REV. 155, 156 (2005) ("[T]he *Batson* peremptory challenge framework is woefully ill-suited to address the problem of race and gender discrimination in jury selection. Current reform proposals are hit or miss, because they do not directly address this source of injustice. Although abolishing the peremptory challenge would be optimal, in the alternative this article recommends several steps that lawyers and judges should take to reduce the impact of unconscious bias on jury selection."). Justice Breyer recently invoked the implicit prejudice view to support his argument for abolition of the peremptory challenge. *See* Miller-El v. Dretke, 545 U.S. 231, 267 (2005) (Breyer, J., concurring) ("And most importantly, at step three, *Batson* asks judges to engage in the awkward, sometime hopeless, task of second-guessing a prosecutor's instinctive judgment—the underlying basis for which may be invisible even to the prosecutor exercising the challenge. In such circumstances, it may be impossible for trial courts to discern if a 'seat-of-the-pants' peremptory challenge reflects a 'seat-of-the-pants' racial stereotype." (citations omitted)).

[13] *See* Reshma M. Saujani, *"The Implicit Association Test": A Measure of Unconscious Racism in Legislative Decision-Making*, 8 MICH. J. RACE & L. 395, 396 (2003) ("The IAT could 'smoke out' illegitimate purposes by demonstrating that the [racially-neutral] classification does not in fact serve its stated purpose."); *id.* at 413 ("The IAT (or psychological testing more generally) should be added to the non-exclusive list enumerated in Arlington Heights for determining racial intent.").

use implicit prejudice research to remake the law. A group of prominent law professors and social psychologists recently joined forces "to use the energy generated by research on unconscious forms of prejudice to understand and challenge the notion of intentionality in the law."[14] The first target is antidiscrimination law's emphasis on intentional discrimination, which these scholars claim "runs afoul of the psychologists' research on implicit prejudice,"[15] but the larger target is "the role of intent in all bodies of law."[16] This empirical updating of the legal significance of the subconscious raises a host of controversial possibilities,[17] including using the results of implicit association tests in Article III confirmations, developing re-conditioning programs for schools and media campaigns to change the unconscious associations we make to the categories of women and minorities, ordering "debiasing screensavers" as an equitable remedy in discrimination lawsuits, and taking measures to reduce the unconscious influence of defendants' Afrocentric facial features in sentencing decisions.[18]

This movement for major changes in the legal understandings of intentionality and discrimination proceeds from the premise that implicit

---

[14] Beth Potier, *Making Case for Concept of 'Implicit Prejudice': Extending the Legal Definition of Discrimination*, HARV. U. GAZETTE (Dec. 16, 2004), *available at* http://www.news.harvard.edu/gazette/2004/12.16/09-prejudice.html (quoting Dr. Mahzarin Banaji). The Russell Sage Foundation is providing funding for an interdisciplinary group of scholars to develop legal applications of the implicit prejudice research. *See The Legal Design of Equality Based on the Science of Ordinary Prejudice*, http://www.russellsage.org/ (type title into search box, then follow hyperlink) (last visited Nov. 14, 2006) (project description). The project is headed by the psychologist Mahzarin Banaji and Katherine Newman, Dean of Social Science at the Radcliffe Institute for Advanced Study. *See id.* Participants in the California Law Review symposium on implicit prejudice research include the psychologists Mahzarin Banaji, Jennifer Eberhardt, Susan Fiske, Anthony Greenwald, John Jost, and Lee Ross and the legal scholars Richard Banks, Gary Blasi, Jerry Kang, and Linda Hamilton Krieger.

[15] Potier, *supra* note 14.

[16] Jerry Kang, *Trojan Horses of Race*, 118 HARV. L. REV. 1489, 1536 (2005).

[17] In 1987, Professor Lawrence authored an influential paper on unconscious discrimination from a psychodynamic perspective, *see* Charles R. Lawrence III, *The Id, the Ego, and Equal Protection: Reckoning with Unconscious Racism*, 39 STAN. L. REV. 317 (1987), but his Freudian language was unfashionable among psychological theorists even before it hit the press and subsequent work has been dominated by cognitive approaches to unconscious discrimination. *See, e.g.*, Martha Chamallas, *Listening to Dr. Fiske: The Easy Case of* Price Waterhouse v. Hopkins, 15 VT. L. REV. 89 (1990); Linda Hamilton Krieger, *The Content of Our Categories: A Cognitive Bias Approach to Discrimination and Equal Employment Opportunity*, 47 STAN. L. REV. 1161 (1995); Mary R. Radford, *Sex Stereotyping and the Promotion of Women to Positions of Power*, 41 HASTINGS L.J. 471 (1990).

[18] *See* Kang, *supra* note 16, at 1536–37.

prejudice research is reliable science with clear real-world implications. Bestowing scientific credibility on this research is essential to the implicit prejudice project, for it insulates the project from charges of value-driven legal prescriptions. If the "science of implicit social cognition"[19] reveals that all individuals exhibit unconscious biases toward women and minorities, then surely the law should take notice if it is committed to equal protection of all persons. Only those suffering from "hypocrisy and self-deception"[20] could cling to a legal regime based on intentional discrimination in the face of scientific research showing the importance of unintentional discrimination.[21] Accordingly, implicit prejudice scholars work hard to claim the mantle of science as they advance their agenda.[22]

It is easy to be overwhelmed by the sheer volume of laboratory studies that implicit prejudice advocates cite, by the moral certitude with which they apply psychological generalizations to the real world, and by the impressive credentials they bring to the courtroom. But this would be a big mistake. On closer inspection, we shall discover that the scientific rhetoric accompanying legal applications of this research is—to use Susan Haack's distinction—more honorific than descriptive:

> So successful have the natural sciences been that the words "science," "scientific," and "scientifically" are often used as generic terms of epistemological praise, meaning vaguely "strong, reliable, good"—as, in television advertisements, actors in white coats urge viewers to get their clothes cleaner with new "scientific" Wizzo.... If "scientific" is used honorifically, it is a tautology that "scientific" equals "reliable"; but this tautology, obviously, is of no help to a judge trying to screen proffered

---

[19] Kang & Banaji, *supra* note 3, at 1064.

[20] *Id.* at 1065.

[21] Kang and Banaji challenge the legal establishment to explain its neglect of this "science": "The law views itself as achieving just, fair, or at least reasonable results. If science reveals that the law is failing to do so because it is predicated on erroneous models of human behavior, then the law must transparently account for the gap instead of ignoring its existence." *Id.*

[22] For example, Professors Kang and Banaji refer to the implicit bias research as "science" or "scientific" at least fifteen times, *see* Kang & Banaji, *supra* note 3, *passim*, yet they offer no discussion of the many challenges to this "science" presently being made within psychology. Professor Kang, in his earlier Harvard Law Review article promoting the implicit prejudice research, calls the work "remarkable science," Kang, *supra* note 16, at 1490, 1592, "jaw-dropping" science, *id.* at 1497, and "the best science known to us," *id.* at 1586. Interestingly, Kang often referred to the work interchangeably as "science" and "social science" in his Harvard article, but Kang and Banaji drop the "social" adjective and treat the work just as "science" in general in their California Law Review article. *See* Kang & Banaji, *supra* note 3, *passim*.

scientific testimony.[23]

We shall document in detail that if implicit prejudice researchers were explicit about the scientific problems underlying key knowledge claims, their extraordinarily ambitious legal-reform project would lose much of its allure. To this end of pulling back the curtain to reveal the messy facts behind neat normative pronouncements,[24] we describe precisely the many ways in which the research behind implicit prejudice claims proves far weaker than its proponents acknowledge. It is worth stressing, though, that our focus is confined to work on implicit prejudice and that our aims are remedial in intent. We do not dispute that the broader implicit social cognition research program has yielded intriguing data bearing on unconscious correlates of judgment under laboratory conditions. Furthermore, we endorse the "behavioral realism" project of subjecting law's behavioral theories to empirical scrutiny,[25] but endorsement of this project does not compel agreement with the conclusions of the implicit prejudice behavioral realists—indeed, acceptance of those conclusions would cast into doubt the scientific bona fides of the project. We simply seek to correct the growing misconceptions that if influential psychologists declare that a measure taps into unconscious prejudice, we are obliged to accept that characterization and, if we accept that characterization, we are further obliged—on pain of being labeled self-deceiving hypocrites—to accept that unconscious prejudice causes legally actionable discriminatory behavior in realistic settings.

We document four major scientific shortcomings of implicit prejudice scholarship:

(a) *Problems of Construct Validity and Metric Meaning: Researchers jump the inferential gun in labeling measures of implicit associations measures of unconscious propensity to discriminate.*

Although rarely acknowledged in law review discussions of implicit

---

[23] Susan Haack, *Trial and Error: The Supreme Court's Philosophy of Science*, 95 AM. J. PUB. HEALTH S66, S68 (2005).

[24] *E.g.*, Kang & Banaji, *supra* note 3, at 1080 ("[W]e need a new model of discrimination for implicit bias—one based on a more accurate model of human cognition and emotion, especially its constraints. This new model must promote proactive structural interventions that minimize harm without relying solely on potential individual litigation." (footnote omitted)).

[25] *See* Linda Hamilton Krieger & Susan T. Fiske, *Behavioral Realism in Employment Discrimination Law: Implicit Bias & Disparate Treatment*, 94 CAL. L. REV. 997, 1000 (explaining that "behavioral realism . . . holds that as judges develop and elaborate substantive legal theories, they should guard against basing their analysis on inaccurate conceptions of relevant, real-world phenomena").

prejudice, sharp dispute exists over what psychological processes the IAT actually measures. IAT proponents claim that the IAT taps into hidden reservoirs of unconscious positive and negative affect toward different social groups, but many studies question this interpretation and indicate that the IAT measures a host of alternative processes that do not involve implicit negative bias toward social groups. For instance, variations in the mere familiarity of the group categories activated by the IAT can lead to scores indistinguishable from those motivated by animus toward those groups;[26] so too can egalitarian empathy for disadvantaged social groups;[27] so too can performance anxiety linked to the fear of being labeled a bigot;[28] so too can mere awareness of cultural stereotypes and depressing socio-demographic facts.[29]

It is unfortunate that so many implicit prejudice scholars fail to discuss evidence casting doubt on the implicit prejudice hypothesis—indeed, some not only ignore this debate but claim a presumption of correctness in their interpretations of this ambiguous evidence by attaching the label "science" to their views.[30] Compounding the problem, these scholars assign social

---

[26] *See* Sachiko Kinoshita & Marie Peek-O'Leary, *Does the Compatibility Effect in the Race Implicit Association Test Reflect Familiarity or Affect?*, 12 PSYCHONOMIC BULL. & REV. 442 (2005). These researchers also call their counter-interpretation "salience asymmetry." *Id.* at 444.

[27] *See* Eric Luis Uhlmann et al., *Are Members of Low Status Groups Perceived as Bad, or Badly Off? Egalitarian Negative Associations and Automatic Prejudice*, 42 J. EXPERIMENTAL SOC. PSYCHOL. 491 (2006).

[28] *See* Cynthia M. Frantz et al., *A Threat in the Computer: The Race Implicit Association Test as a Stereotype Threat Experience*, 30 PERSONALITY & SOC. PSYCHOL. BULL. 1611 (2004).

[29] *See* Hal R. Arkes & Philip E. Tetlock, *Attributions of Implicit Prejudice, or "Would Jesse Jackson 'Fail' the Implicit Association Test?,"* 15 PSYCHOL. INQUIRY 257, 268–74 (2004); Philip E. Tetlock & Hal R. Arkes, *The Implicit Prejudice Exchange: Islands of Consensus in a Sea of Controversy*, 15 PSYCHOL. INQUIRY 311, 316 (2004). One of the few points of agreement in a recent exchange involving one of the co-authors of the present paper and proponents of the IAT is that, within the implicit prejudice research program, it is perfectly possible for a citizen to be a Bayesian bigot—a bigot by virtue of simply rationally processing available information in his or her society about depressing patterns of covariation between race and socioeconomic outcomes. *See* Mahzarin R. Banaji et al., Commentary, *No Place for Nostalgia in Science: A Response to Arkes and Tetlock*, 15 PSYCHOL. INQUIRY 279, 283–85 (2004); Tetlock & Arkes, *supra*, at 319.

[30] *See* Kang & Banaji, *supra* note 3, at 1076 ("First, responding to discrimination should be a constitutionally compelling interest regardless of whether explicit or implicit bias actuates the discrimination. Those who argue otherwise must confront the science that demonstrates the existence and real-world consequences of implicit bias. Given this evidence, they bear the burden to show why these harms . . . should be categorically

meaning to scores on the IAT that lack empirical justification. The IAT is an arbitrary metric that sorts people along a dimension—reaction time—that looks objective but lacks any objective connections to legally actionable behavior. Thus, even if we grant that the IAT is a valid measure of implicit associations between group categories and evaluative attitudes, IAT scores remain meaningless until empirical studies link specific ranges of scores to specific acts that objectively (or consensually) represent discrimination. To date, this essential mapping task has not been undertaken. Yet, when claims about the meaning of IAT studies are unpacked, they invariably imply that particular scores on this arbitrary metric hold non-arbitrary social meaning (e.g., IAT studies show *strong* implicit bias *against* the elderly or high levels of implicit *racism*).

Scholars who blur this distinction violate a canonical scientific norm: the injunction to separate factual from value judgments. Attributions of prejudice inevitably rest on complex amalgams of factual and value assumptions, and it is a mistake to suppose that, just because a select group of social psychologists and law professors—with a self-declared agenda to transform American law—announce the discovery of a new form of prejudice, the rest of society is obliged to defer to their judgment. This is particularly true when these academics have set their threshold for declaring people prejudiced so low that the vast majority of the American population qualifies as bigoted by blinking more frequently in the presence of a minority or by associating elderly persons with positive traits milliseconds slower than they associate young persons with positive traits.[31] These social psychologists and legal scholars are claiming, in effect, not only scientific expertise on factors that sway human judgment but also the moral authority to determine where society should draw the line between extremely subtle forms of "prejudice" and behaviors that warrant no censure.

(b) *Problems of Internal Validity: Researchers ignore alternative explanations for alleged discriminatory behavior that conflict with the implicit-prejudice hypothesis.*

Implicit prejudice researchers rely entirely on correlational evidence to find a relationship between implicit prejudice (as supposedly measured by the IAT) and discrimination (as broadly defined by these researchers to include awkward social interactions), and these correlational studies rarely

---

disregarded simply because their causes operate beneath our self-awareness." (footnotes omitted)).

[31] For a full description of the methods used in implicit prejudice research, see *infra* Part II.B. For instructive comparisons of how sharp a departure this approach represents from traditional prejudice research, as well as from common sense, see GORDON W. ALLPORT, THE NATURE OF PREJUDICE (1954); RUPERT BROWN, PREJUDICE: ITS SOCIAL PSYCHOLOGY (1995); Peter Suedfeld, Commentary, *Racism in the Brain; or is it Racism on the Brain?*, 15 PSYCHOL. INQUIRY 298 (2004).

control for a variety of confounding factors that could explain the pattern of results without assuming implicit prejudice or stereotypes at work. Indeed, these studies fail to control for variables that, if causing the results reported, would radically transform the moral meaning of test scores, raising the possibility that, far from feeling deep-seated animus toward minorities, subjects feel guilty about America's history of discrimination—feelings that lead them to act in ways that are signs not of discrimination but rather discomfort and shame.

(c) *Problems of Statistical-Conclusion Validity: The IAT has serious psychometric flaws and an alarmingly high false alarm rate.*

Simple Bayesian analysis reveals that the modest correlation coefficients that implicit prejudice researchers invoke to support their claims about the pervasiveness of prejudice, coupled with the high failure rate on the test, are bound to lead to many false accusations of implicit bigotry. This result is also not surprising in view of the complex determinants of performance on the IAT. IAT scores depend on reaction times to shifting combinations of stimuli (e.g., Black + unpleasant, White + pleasant), with quicker reactions taken as evidence of stronger associations between the stimuli, and these presumed associations then treated as evidence of implicit biases for or against the different groups represented by the stimuli. Psychometric studies have shown that a host of factors other than association strength can affect reaction time (e.g., cognitive flexibility, asymmetries in stimuli familiarity, and evaluation apprehension), yet IAT researchers assume that different reaction times signify only different attitudes and biases toward the stimuli.

(d) *Problems of External Validity: Researchers suspend disbelief in judging the real-world implications of laboratory results on implicit prejudice.*

Even if there were no doubt that implicit measures of prejudice capture what they purport to measure and reliably predicted discriminatory behavior in experimental studies, those eager to import this research into the law still must establish that the correlations between IAT scores and discriminatory conduct found in artificial laboratory settings reliably predict behavior in real-world settings that often have institutionalized layers of safeguards against the expression of prejudice.[32] Empirical tests validate these

---

[32] *See, e.g.*, Winfred Arthur, Jr. & Dennis Doverspike, *Achieving Diversity and Reducing Discrimination in the Workplace Through Human Resource Management Practices: Implications of Research and Theory for Staffing, Training, and Rewarding Performance*, *in* DISCRIMINATION AT WORK: THE PSYCHOLOGICAL AND ORGANIZATIONAL BASES 305 (Robert L. Dipboye & Adrienne Colella eds., 2005) (reviewing human resources practices aimed at reducing discrimination); Sturm, *supra* note 1, at 489–522 (discussing organizational innovations to fight discrimination); U.S. Equal Employment

safeguards as effective antidiscrimination measures, and, given the weak relationships between IAT scores and discriminatory behavior under even "ideal" laboratory conditions for eliciting discrimination, there is no reason to believe that these safeguards will not be effective against discrimination motivated by implicit biases.

In Part III, we flesh out these criticisms. But first, in Part II, we provide a fuller description of psychological research on implicit prejudice and legal scholarship that relies on this research. We then detail the numerous problems of scientific validity that plague the research program and discuss the conceptual confusions within this body of work that undercut its credence both as legislative authority and litigation evidence. We conclude by addressing the severe fact-value problems inherent in implicit prejudice research and discussing how implicit extra-scientific values can blind research communities to scientific questions that cast doubt on both factual claims and policy recommendations.

II. SOCIAL SCIENCE AND THE ACADEMIC MOVE FROM CONSCIOUS TO UNCONSCIOUS DISCRIMINATION

A. *Terminology*

To avoid unnecessary confusion, we pause to specify our intended meanings for a variety of terms employed in legal and social scientific discussions of discrimination. Although we provide common definitions of the relevant psychological constructs and legal concepts, we claim no special authority for our definitions. Most importantly, none of our criticisms of the implicit prejudice research depends on the reader's acceptance of our definitions. In our explication of the key psychological constructs, we follow the lead of others who distinguish among the cognitive, affective, and behavioral components of discrimination.[33]

Opportunity Commission, Best Practices of Private Sector Employees, http://www.eeoc.gov/abouteeoc/task_reports/prac2.html (last visited Sept. 8, 2006).

[33] *See* John Duckitt, *Prejudice and Intergroup Hostility*, *in* OXFORD HANDBOOK OF POLITICAL PSYCHOLOGY 559, 559 (David O. Sears et al. eds., 2003) ("Social psychologists have distinguished three distinct components of prejudice, or ways in which negative intergroup attitudes can be expressed. These are negative stereotypes (cognitive component), negative feelings (affective component), and negative behavioral inclinations (behavioral component) toward outgroups."); Susan T. Fiske, *Stereotyping, Prejudice, and Discrimination*, *in* 2 HANDBOOK OF SOCIAL PSYCHOLOGY 357, 357 (Daniel T. Gilbert, Susan T. Fiske & Gardner Lindzey eds., 4th ed. 1998) ("Following one traditional division of attitudes, stereotyping is taken as the most cognitive component, prejudice as the most affective component, and discrimination as the most behavioral component of category-based reactions . . . ." (citations omitted)).

First, "bias" refers to systematic variation in judgmental tendencies elicited by some attribute or property of a stimulus, such as a person's membership in a particular group. For instance, memory for faces exhibits a same-race bias: eyewitnesses more accurately recall same-race faces than cross-race faces.[34]

Bias may be implicit, in which case people do not recognize the biasing influence of a stimulus on their judgment at the time of its operation, or explicit, in which case people do recognize the biasing influence.[35] An implicit cognitive bias, so defined, operates automatically, because it falls at the time of judgment beyond conscious awareness or intentional control.[36] However, the antecedent conditions must be right for an implicit bias to be automatically activated and then, in turn, to influence outward behavior. And an implicit bias can suddenly become explicit if the social context alerts people to the direction and magnitude of the bias, thereby making self-correction possible.[37] Failure to appreciate these triggering and moderating conditions can lead to alarming but ultimately groundless claims about the prevalence and power of implicit biases.[38]

Bias may be elicited by an objective attribute of a stimulus, such as the physiognomy of a human face in the case of the same-race bias for facial memory, or by inferences about the attributes of a stimulus drawn from related stimuli, such as when group stereotypes influence judgments about

---

[34] *See* Christian Meissner & John C. Brigham, *Thirty Years of Investigating the Own-Race Bias in Memory for Faces: A Meta-Analytic Review*, 7 PSYCHOL. PUB. POL'Y & L. 3 (2001).

[35] We mean here to track Greenwald's two definitions of unconscious: (1) processes or stimuli outside of attention and (2) mental processes or events that escape accurate introspection. *See* Anthony G. Greenwald, *New Look 3: Unconscious Cognition Reclaimed*, 47 AM. PSYCHOL. 766, 767 (1992).

[36] *See* John A. Bargh & Tanya L. Chartrand, *The Unbearable Automaticity of Being*, 54 AM. PSYCHOL. 462, 463–64 (1999) (contrasting conscious and automatic mental processes); Nilanjana Dasgupta, *Implicit Ingroup Favoritism, Outgroup Favoritism, and Their Behavioral Manifestations*, 17 SOC. JUST. RES. 143, 144 n.2 (2004) ("Typically, psychological responses measured in research studies have been called 'implicit,' 'automatic,' or 'nonconscious' to the extent that at least one of the primary criteria—lack of awareness, intention, or control—has been operational.").

[37] *See, e.g.*, Richard E. Petty & Duane T. Wegener, *Attitude Change: Multiple Roles for Persuasion Variables*, *in* 1 HANDBOOK OF SOCIAL PSYCHOLOGY 323, 330–31 (Daniel T. Gilbert, Susan T. Fiske & Gardner Lindzey eds., 4th ed. 1998); Philip E. Tetlock, *Cognitive Biases and Organizational Correctives: Do Both Disease and Cure Depend on the Political Beholder?*, 45 ADMIN. SCI. Q. 293 (2000).

[38] *See infra* Part III.

individuals (e.g., this Republican will act like other Republicans).[39] Although reliance on stereotypes can lead to biased beliefs and biased attributions about members of social groups,[40] such reliance can also sometimes satisfy technical definitions of individual rationality (that is, stereotypes may have predictive value, and stereotype-driven bias should not be conflated with irrationality or animus).[41]

---

[39] *See* Robert L. Dipboye & Adrienne Colella, *An Introduction*, *in* DISCRIMINATION AT WORK, *supra* note 32, at 1 ("*Stereotyping* is used to refer to the cognitive biases against outgroup members and includes not only attributions of traits to members of these groups but also beliefs about these individuals."); Eliot R. Smith, *Mental Representation and Memory*, *in* 1 HANDBOOK OF SOCIAL PSYCHOLOGY, *supra* note 37, at 400 ("Stereotypes have often been conceptualized as associative links between a node representing a social group and various traits and/or evaluations."). The traditional definition of stereotype emphasizes assimilation effects: the individual is assimilated to the group. Recent work emphasizes that stereotypes may serve more generally as standards for comparing individuals to groups; under this view, stereotypes can also lead to contrast effects that generate counter-stereotypical responses to an individual. *See* Monica Biernat, *Toward a Broader View of Social Stereotyping*, 58 AM. PSYCHOL. 1019 (2003).

[40] *See, e.g.*, Celina M. Chatman & William von Hippel, *Attributional Mediation of In-Group Bias*, 37 J. EXPERIMENTAL SOC. PSYCHOL. 267, 271 (2001) ("[W]e have shown that Blacks and Whites are subject to in-group biases in their attributions for the behavior of other Blacks and Whites and these attributions at least partially account for biased evaluations of in-group and out-group members."); Patricia G. Devine, *Stereotypes and Prejudice: Their Automatic and Controlled Components*, 56 J. PERSONALITY & SOC. PSYCHOL. 5, 5 (1989) ("[M]any classic and contemporary theorists have suggested that prejudice is an inevitable consequence of ordinary categorization (stereotyping) processes." (citations omitted)).

[41] The belief that stereotypes generally lead to erroneous judgments about targets may itself be an erroneous stereotype about stereotypes, for research suggests that surprisingly often "group stereotypes and perceptions of members of stereotyped groups can be quite accurate," Clark R. McCauley et al., *Stereotype Accuracy: Toward Appreciating Group Differences*, *in* STEREOTYPE ACCURACY 293, 297 (Yueh-Ting Lee et al. eds., 1995), and stereotype-driven responding can satisfy technical standards of rationality when the stereotype provides better information than individualized judgments. *See* Amy Farmer & Dek Terrell, *Crime Versus Justice: Is There a Trade-Off?*, 44 J.L. & ECON. 345, 345–46 (2001) ("Models of statistical discrimination show that imperfect information regarding an individual's characteristics may lead people to use group membership to assist in decision making."); Roland G. Fryer, Jr. & Steven D. Levitt, *The Causes and Consequences of Distinctively Black Names*, 119 Q.J. ECON. 767, 801 (2004) ("[W]hile we cannot rule out animus on the part of employers, we find evidence supporting a potential productivity-related statistical discrimination motive for employers to base interview decisions on first names."); *see also* Fiske, *supra* note 33, at 375 ("Categorical reactions persist in part because they are cognitively useful. They also persist because they are socially useful."). For example, using the race of a person approaching on a dark street in a high crime area to predict that person's criminal

Second, "prejudice" refers to a systematic affective or evaluative response to a social group and its members.[42] Traditionally, prejudice implied a negative attitude toward a particular group, but definitional fashions change and some researchers have dropped animus as a necessary feature of prejudice.[43] Thus, "benevolent sexism," a protective, paternalistic attitude toward women, may be a form of prejudice under this revised view.[44] Prejudice, as we define it, is a special type of attitude reserved for groups,[45] and prejudicial attitudes may be implicit or explicit in the same

---

propensity may be rational if one's race-based stereotypes reflect true differences in base rates of criminality across racial groups. Nevertheless, lawmakers may choose to forbid use of such base rate information on grounds that discrimination contributed to these base rates or that the law favors particularized judgments about people, or that the costs of using stereotypes outweigh the benefits. *See* Jody D. Armour, *Race Ipsa Loquitur: Of Reasonable Racists, Intelligent Bayesians, and Involuntary Negrophobes*, 46 STAN. L. REV. 781, 790–96 (1994). A sizable segment of the citizenry directs moral outrage at those who use forbidden (race-charged) base rates and display signs of guilt and desire to engage in moral cleansing when they discover that they have inadvertently used such base rates by relying on predictor variables that later prove to be correlated with, and thus contaminated by, the base rates. *See* Philip E. Tetlock et al., *The Psychology of the Unthinkable: Taboo Trade-Offs, Forbidden Base Rates, and Heretical Counterfactuals*, 78 J. PERSONALITY & SOC. PSYCHOL. 853, 863–64 (2000).

[42] *Cf.* Dipboye & Colella, *supra* note 39, at 1 ("We refer to *prejudice* as the attitudinal and especially the affective biases that exist with regard to members of groups other than those to which one belongs."); John F. Dovidio, *On the Nature of Contemporary Prejudice: The Third Wave*, 57 J. SOC. ISSUES 829, 829 (2001) ("Prejudice is commonly defined as an unfair negative attitude toward a social group or a person perceived to be a member of that group.").

[43] *See* Christian S. Crandall & Amy Eshleman, *A Justification-Suppression Model of the Expression and Experience of Prejudice*, 129 PSYCHOL. BULL. 414, 414–15 (2003) (noting that positive prejudice may exist but that negative prejudice remains dominant and more problematic); *see also* JOHN DUCKITT, THE SOCIAL PSYCHOLOGY OF PREJUDICE 17 (1992) ("Because the idea of prejudice as a bad and unjustified attitude has always been an essentially subjective value judgment, prejudice has never actually been operationalized and measured in that way."). In addition, the current view is that prejudice need not arise from false or inaccurate beliefs about a social group. *See* Crandall & Eshleman, *supra*, at 414; *see also* BROWN, *supra* note 31, at 6–8.

[44] *See* Peter Glick & Susan T. Fiske, *An Ambivalent Alliance: Hostile and Benevolent Sexism as Complementary Justifications for Gender Inequality*, 56 AM. PSYCHOL. 109, 116 (2001) ("Although sexist antipathy is the most obvious form of prejudice against women, our evidence suggests that sexist benevolence may also play a significant role in justifying gender inequality.").

[45] We define attitude as a "psychological tendency that is expressed by evaluating a particular entity with some degree of favor or disfavor." Alice H. Eagly & Shelly Chaiken, *Attitude Structure and Function*, *in* 1 HANDBOOK OF SOCIAL PSYCHOLOGY, *supra* note 37, at 269; *accord* Anthony G. Greenwald et al., *A Unified Theory of Implicit Attitudes, Stereotypes, Self-Esteem, and Self-Concept*, 109 PSYCHOL. REV. 3, 5 (2002) ("An *attitude* is the association of a social object or social group concept with a valence

way that cognitive bias may be implicit or explicit.

Third, "discrimination" refers to the behavioral consequences of a group bias (typically in the form of a group stereotype) or a prejudice toward a particular group.[46] Under this expansive *psychological* definition, any behavioral consequence of group bias or prejudice counts as discrimination. It is worth noting though that the current *legal* definition of discrimination is much narrower.[47] Courts typically require that the adverse action against the plaintiff rise above what Dean White calls "de minimis" discrimination.[48]

---

attribute concept."); *see also* Eagly & Chaiken, *supra*, at 270 ("[A]ttitudes toward minority groups are often called *prejudice*, especially if these attitudes are negative.").

[46] *Cf.* Dipboye & Colella, *supra* note 39, at 2 ("*Discrimination* refers to the unfair behavioral biases demonstrated against [outgroup members]."); Fiske, *supra* note 33, at 374–75 ("In all likelihood, there are two kinds of discrimination to document . . . . [These include] 'hot discrimination,' based on disgust, resentment, hostility, and anger . . . [and] 'cold discrimination,' based on stereotypes of an outgroup's interests, knowledge, and motivations.").

[47] In addition to law's narrower definition of discrimination, constitutional and statutory limits on the scope of protection provided by antidiscrimination law further shrink the class of acts that qualify as illegal discrimination compared to the class of acts that qualify as psychological discrimination. *See, e.g.*, 42 U.S.C. § 2000e(b) (2000) (defining covered employers to be persons "engaged in an industry affecting commerce who has fifteen or more employees for each working day in each of twenty or more calendar weeks in the current or preceding calendar year").

We recognize that legal conceptions of discrimination vary, *see generally* George Rutherglen, *Discrimination and Its Discontents*, 81 VA. L. REV. 117 (1995), but we employ here fairly standard definitions from employment law. Arguably, under current law, disparate treatment claims can encompass acts of discrimination caused by conscious and unconscious influences of a target's protected characteristic(s). *See, e.g.*, Michael Selmi, *Proving Intentional Discrimination: The Reality of Supreme Court Rhetoric*, 86 GEO. L.J. 279, 294 (1997) ("In defining intentional discrimination, the question is not what the particular decisionmaker subjectively intended, but whether the record allows for an inference that an impermissible factor such as race served as the impetus for the challenged action." (footnote omitted)); Amy L. Wax, *Discrimination as Accident*, 74 IND. L.J. 1129, 1138–39 (1999) (Disparate treatment can "narrowly denote a form of scienter—an actor's conscious awareness of his reasons for acting. But it can also be used more broadly to refer to a causal link between a mental influence . . . and the outcome of a decision." (footnote omitted)).

[48] *See, e.g.*, Tristin K. Green, *Workplace Culture and Discrimination*, 93 CAL. L. REV. 623, 655 (2005) ("To prevail on a claim of disparate treatment, a plaintiff typically must identify a particular decision maker who has taken a 'materially adverse' employment action against her, and she must prove intent to discriminate, frequently construed as conscious bias or animus, on the part of the decision maker." (footnotes omitted)). For a critical analysis of the "de minimis discrimination" concept as it has developed within employment discrimination law, see Rebecca Hanner White, *De Minimis Discrimination*, 47 EMORY L.J. 1121 (1998). White argues that, in cases of direct employer liability, "[e]mployer policies that distinguish between and among workers based on race or sex are within the scope of [Title VII], regardless of how trivial or de

Thus, many subtle acts by managers or co-workers that psychologists would label discriminatory do not rise to the level of illegal discrimination unless accompanied by some tangible effect or unless they cumulatively create a hostile work environment.[49]

---

minimis such discrimination may seem. But when vicarious, not direct, liability is at issue, there is a de minimis threshold." *Id.* at 1191. As Dean White notes, however, many courts do not follow her prescription and, indeed, the trend seems to be toward expansion of the *de minimis* exclusion. *See id.* at 1122, 1143; *see also* Rosalie Berger Levinson, *Parsing the Meaning of "Adverse Employment Action" in Title VII Disparate Treatment, Sexual Harassment, and Retaliation Claims: What Should Be Actionable Wrongdoing?*, 56 OKLA. L. REV. 623, 623 n.3 (2003) ("The cases discussed in this Article suggest that courts, even within the so-called 'expansive' and 'intermediate' circuits, are issuing decisions that reflect a restrictive view as to what harms are actionable both in the context of retaliation and disparate treatment claims."). For examples from the case law, see Boone v. Goldin, 178 F.3d 253, 256 (4th Cir. 1999) ("Congress did not intend Title VII to provide redress for trivial discomforts endemic to employment."); Holt v. Morgan, 79 F. App'x 139, 141 (6th Cir. 2003) ("A performance evaluation that is lower than an employee feels is warranted is not an adverse employment action sufficient to state a claim of discrimination."); Hopkins v. Women's Div., Gen. Bd. of Ministry, 98 F. App'x 8, 9 (D.C. Cir. 2004) ("Neither declining to discipline a support-staff member nor depriving someone of office supplies is a 'tangible employment action evidenced by firing, failing to promote, a considerable change in benefits, or reassignment with significantly different responsibilities.'" (citation omitted)); *see also* Keeton v. Flying J, Inc., 429 F.3d 259, 263 n.1 (6th Cir. 2005) ("The terms 'tangible employment action' and 'adverse employment action' are interchangeable.").

Subjecting a plaintiff to a hostile work environment constitutes discrimination even if it does not result in other, tangible changes in employment conditions. *See* Meritor Sav. Bank, FSB v. Vinson, 477 U.S. 57, 67 (1986) (holding that severe or pervasive sexual harassment may constitute an actionable change in employment conditions under Title VII). However, employers can invoke the affirmative defense set out in *Ellerth* if the harassment does not result in tangible employment action against the harassed employee. *See* Burlington Indus., Inc. v. Ellerth, 524 U.S. 742, 765 (1998). And not all harassment creates an illegal, hostile work environment. *See* Faragher v. City of Boca Raton, 524 U.S. 775, 788 (1998) ("A recurring point in [our] opinions is that 'simple teasing,' offhand comments, and isolated incidents (unless extremely serious) will not amount to discriminatory changes in the 'terms and conditions of employment.'" (citation omitted)).

[49] *See* Dolly Chugh, *Societal and Managerial Implications of Implicit Social Cognition: Why Milliseconds Matter*, 17 SOC. JUST. RES. 203, 209 (2004) ("[A]n almost uncountable number of micro-behaviors may affect the actual fairness of how an individual is treated after being granted the opportunity to be employed, or be schooled, or be treated, or be tried."). To be sure, some legal scholars have argued for a broadening of the legal definition of discrimination, *see, e.g.*, Green, *supra* note 48, at 684 ("I have purposefully pushed our conception of law beyond legal rights to explore alternatives for combating some of the more subtle, ongoing forms of discrimination that operate alongside those that are traditionally recognized."), but courts remain "reluctant to define subtle discrimination as unlawful discrimination . . . ." Michael Selmi, *Subtle Discrimination: A Matter of Perspective Rather than Intent*, 34 COLUM. HUM. RTS. L. REV. 657, 659 (2003). *Cf.* Bagenstos, *supra* note 1, at 17–18 ("These features of disparate

In our view, bias, prejudice, and discrimination are best treated as distinct psychological constructs: one may process information about a member of an out-group in a cognitively biased manner, but this biased processing need not definitively determine one's personal attitude or behavior toward the target. For instance, sex stereotypes may lead an employer to assume that a woman is more likely than a man to have an interest in child-rearing, but this belief bias does not necessarily have positive or negative attitudinal or behavioral implications. The employer might conclude that women are more emotionally supportive and better team players or that women are more easily distracted by family demands. Likewise, prejudice toward a group may operate independently of one's stereotype about that group.[50] Finally, a target's membership in a particular group may trigger cognitive bias and prejudice but not discriminatory behavior if other psychological processes or situational conditions override the bias and prejudice. Of course, although these constructs can be logically separated, they remain empirically connected: stereotypes can, and sometimes do, reinforce prejudicial attitudes; prejudices can, and sometimes do, prevent interactions with out-group members that could falsify stereotypic beliefs and thereby facilitate biased assimilation of new information to existing stereotypes; and stereotypes and prejudices can, and sometimes do, have discriminatory behavioral consequences.[51] These connections are, however, far from automatic. Much hinges, as we shall see, on the organizational context within which the actors are embedded.[52]

---

impact doctrine make it a poor tool for addressing discrimination that does its work through an accumulation of small moments of perception and evaluation." (footnote omitted)).

[50] *See* Fiske, *supra* note 33, at 372–74 (discussing the role of intergroup relations, threat, and emotion in prejudice).

[51] *See* Brad J. Bushman & Angelica M. Bonacci, *You've Got Mail: Using E-Mail to Examine the Effect of Prejudiced Attitudes on Discrimination Against Arabs*, 40 J. EXPERIMENTAL SOC. PSYCHOL. 753, 754 (2004) ("Once social categorization occurs, prejudice, and discrimination are more likely to follow."); Scott Plous, *The Psychology of Prejudice, Stereotyping, and Discrimination: An Overview*, *in* UNDERSTANDING PREJUDICE AND DISCRIMINATION 3, 5 (Scott Plous ed., 2003) ("[P]rejudice, stereotyping, and discrimination are distinct from one another, even though in daily life they often occur together."); Jeffrey W. Sherman et al., *Prejudice and Stereotype Maintenance Processes: Attention, Attribution, and Individuation*, 89 J. PERSONALITY & SOC. PSYCHOL. 607, 607 (2005) ("[A] central theme in social psychological theory has been that stereotyping promotes prejudice and that prejudice reduction depends on stereotype change.").

[52] *See infra* Part III.D.

B. *The Implicit Prejudice Research Program*

Two themes dominate the history of social psychological research on intergroup conflict: (a) continual adjustment of measures and standards for assessing the prevalence of intergroup hostility[53] and (b) constant revision of the psychological explanations for the sources of intergroup hostility. For the second theme, the focus has shifted with prevailing intellectual fashions from psychodynamic theories to social-identity theories to cognitive-bias theories to the recent fascination with reaction-time-based associationist theories.[54] The implicit prejudice research program falls into the reaction-time-based associationist theoretical camp.

Following passage of civil rights legislation in the 1960s, overt expressions of racism declined significantly, but large disparities in group outcomes persisted.[55] This disjunction led many racism researchers to suspect that intergroup hostility persisted but had begun manifesting itself in more disguised, socially acceptable, forms.[56] Accordingly, these

---

[53] *See* Markus Brauer et al., *Implicit and Explicit Components of Prejudice*, 4 REV. GEN. PSYCHOL. 79, 79 (2000) ("For more than 70 years, social psychologists have been concerned with the measurement of prejudice toward out-groups . . . . Although the basic goal has remained the same, measurement techniques have changed considerably over the years."); *see also* Suedfeld, *supra* note 31, at 298–300 (providing a historical summary of measurement techniques used in prejudice research).

[54] *See* DUCKITT, *supra* note 43, at 43 ("Many theories have been proposed to explain the causation of prejudice . . . . Unfortunately, these developments have not brought much clarification to the overall question of the causation of prejudice. The list of possible causes, and the complexity of the problem, seems to have increased rather than decreased."); *see also* Susan T. Fiske, *Intent and Ordinary Bias: Unintended Thought and Social Motivation Create Casual Prejudice*, 17 SOC. JUST. RES. 117, 118–23 (2004) (discussing the evolution of psychological explanations for prejudice).

[55] *See* John F. Dovidio & Samuel L. Gaertner, *Aversive Racism and Selection Decisions: 1989 and 1999*, 11 PSYCHOL. SCI. 315, 315 (2000) ("In part because of changing norms and the Civil Rights Act and other legislative interventions that have made discrimination not simply immoral but also illegal, overt expressions of prejudice have declined significantly over the past 35 years." (citation omitted)).

[56] *See id.* ("Discrimination, however, continues to exist and affect the lives of people of color and women in significant ways. What accounts for this discrepancy? One possibility is that it represents a change in the nature of racial prejudice. Contemporary forms of prejudice may be less conscious and more subtle than the overt, traditional form." (citations omitted)); David O. Sears & P.J. Henry, *Over Thirty Years Later: A Contemporary Look at Symbolic Racism*, 37 ADVANCES EXPERIMENTAL SOC. PSYCHOL. 95 (2005) ("This is the problem that has animated our own research agenda: how to understand White's continuing resistance to efforts to increase racial equality despite much evidence that in some measurable ways their racial attitudes have become substantially liberalized.").

psychologists developed less obtrusive methods for measuring racism and reconsidered the psychological mechanisms that lead to discrimination.[57]

The first wave of research into new forms of racism employed indirect measures that inferred racism from the policies and practices that people supported, on the theory that subtle racism was most likely to be expressed in situations where a nonracial justification was available.[58] Much of this new work using less obtrusive methods was taken to confirm the researchers' suspicions that racism remained widespread, if less obvious in its expression, and that racism was "rooted in *normal*, often adaptive, psychological processes,"[59] rather than abnormal psychological processes.

This first wave of research into modern racism proved controversial, however, because many of these indirect measures of racism equated endorsement of a conservative ideology with racist opposition to various liberal policies.[60] For instance, opposition to affirmative action and school busing for integration purposes were originally taken as evidence of "symbolic racism," without first eliminating conservative, non-racist principles and values as the true motivation.[61] Also, because these indirect

---

[57] *See* BROWN, *supra* note 31, at 217 ("The observation that levels of overt prejudice were falling whilst other forms of discrimination were continuing has stimulated a number of new conceptualizations of prejudice over the past 20 years.").

[58] *See* Dovidio, *supra* note 42, at 835 ("Because aversive racists consciously endorse egalitarian values, they will not discriminate directly and openly in ways that can be attributed to racism; however, because of their negative feelings they will discriminate, often unintentionally, when their behavior can be justified on the basis of some factor other than race (e.g., questionable qualifications for a position).").

[59] *Id.* at 834.

[60] For a full discussion of this problem and others with indirect measures of racism, see generally Paul M. Sniderman & Philip E. Tetlock, *Symbolic Racism: Problems of Motive Attribution in Political Analysis*, 42 J. SOC. ISSUES 129 (1986) [hereinafter Sniderman & Tetlock, *Symbolic Racism*]; Paul M. Sniderman & Philip E. Tetlock, *Reflections on American Racism*, 42 J. SOC. ISSUES 173 (1986); Philip E. Tetlock, *Political Psychology or Politicized Psychology: Is the Road to Scientific Hell Paved with Good Moral Intentions?*, 15 POL. PSYCHOL. 509 (1994). *But see generally* Sears & Henry, *supra* note 56 (responding to these criticisms).

[61] In partial recognition of the validity of this criticism, items confounding possible conservative opposition to government action with racism have been dropped from the latest version of the symbolic racism scale. Sears & Henry, *supra* note 56, at 115. Even these revised scales are still bedeviled, however, by the "Bill Cosby" problem: one can get high scores by virtue of endorsing the well-known African-American actor's concerns about the decline of traditional family and work values in Black communities. For example, strong agreement with the first two items in the scale ((1)"It's really a matter of some people not trying hard enough; if blacks would only try harder they could be just as well off as whites[,]" and (2) "Irish, Italian, Jewish and many other minorities overcame prejudice and worked their way up. Blacks should do the same") is scored as evidence of

measures still required public acts or self-reports of conscious attitudes and beliefs with racial overtones, some remained skeptical that these new techniques could accurately measure underlying racial hostilities—either because of impression management or lack of conscious access to the thought processes that may drive intergroup hostility.[62] Accordingly, researchers sought ever more ingenious ways of measuring racism that would avoid the problems of social-desirability contamination and unreliable introspection.[63]

---

symbolic racism. *See* David O. Sears & P.J. Henry, Symbolic Racism Scale, http://condor.depaul.edu/~phenry1/SR2Kinstructions.htm (providing the Symbolic Racism 2000 scale and scoring instructions) (last visited Sept. 28, 2006); *see also* Vincent L. Hutchings & Nicholas A. Valentino, *The Centrality of Race in American Politics*, 6–7 ANN. REV. POL. SCI. 383, 391–92 (2004). When prominent civil rights advocates are logically fated to fail alleged tests of racism, it is not unreasonable to suspect the problem may lie with the tests.

[62] *See* Brauer et al., *supra* note 53, at 80 ("Even these more subtle measures, however, seemed to many researchers too vulnerable to self-presentation."); T. ANDREW POEHLMAN ET AL., UNDERSTANDING THE IMPLICIT ASSOCIATION TEST: III. META-ANALYSIS OF PREDICTIVE VALIDITY 6 (2005) (manuscript on file with authors) ("Many researchers have argued that subjects' desire to provide socially desirable responses results in inaccurate answers on self-report questionnaires . . . . In contrast, because IAT measures resist faking, they may be able to predict criterion measures equally well in socially sensitive and non-sensitive domains." (citations omitted)).

[63] Another popular approach in the first wave of research on "modern racism" was the "bogus pipeline." In these studies, subjects are led to believe that they are to be hooked up to a machine that can take physiological measurements that will reveal their true attitudes toward minorities. In the guise of providing responses that can be used to validate the machine's operation, subjects are then given the key measures of racial attitudes, on the theory that subjects will have a motive to be truthful or their deception will be revealed by the machine's measurements (hence, the ruse was thought to provide a "bogus pipeline" into true thoughts). The responses of these treatment subjects are then compared to the responses of a control group, with any negative difference in racial attitudes in the treatment group being taken as evidence of racism. *See* Edward E. Jones & Harold Sigall, *The Bogus Pipeline: A New Paradigm for Measuring Affect and Attitude*, 76 PSYCHOL. BULL. 349 (1971) (first presenting the methodology). For a review and evaluation of this line of research, see Neal J. Roese & David W. Jamieson, *Twenty Years of Bogus Pipeline Research: A Critical Review and Meta-Analysis*, 114 PSYCHOL. BULL. 363 (1993). Fazio and colleagues later went so far as to declare their affective-priming measure to be "a bona fide pipeline" into subjects' minds superior to explicit measures of attitudes—a claim that Fazio subsequently amended and admitted had led to confusion. *Compare* Russell H. Fazio et al., *Variability in Automatic Activation as an Unobtrusive Measure of Racial Attitudes: A Bona Fide Pipeline?*, 69 J. PERSONALITY & SOC. PSYCHOL. 1013, 1025 (1995) ("The priming procedure appears to provide a bona fide pipeline for attitude measurement."), *with* Fazio & Olson, *supra* note 5, at 304 n.2 (noting that the claims of Fazio et al. caused confusion about whether implicit or explicit measures tap into "real attitudes" and stating that by referring to implicit measures as

These efforts culminated in the second, and most recent, wave of research into modern racism using sophisticated techniques designed to reveal the subconscious processes that drive racism. These methods seek to overcome the subject's motivation to obfuscate and inability to report on mental states by gaining access to the inner workings of the mind via examination of uncontrolled by-products of mental processes, such as physiological indicators, micro-movements of the face, and reaction times.[64] The two most popular methods of "implicit measurement" of intergroup attitudes are affective-priming procedures and tests of implicit association,[65] with the latter now the dominant method for studying implicit stereotypes and prejudice.[66] Both methods are designed to examine natural associations between groups and positively- and negatively-valenced terms to determine implicit attitudes.[67]

In the affective-priming approach, subjects are exposed to an attitude object (the "prime") and then to an evaluative adjective (the "target") that

---

"bona fide" Fazio et al. meant "to indicate that any automatic attitude activation occurs farther upstream than the overt response to an explicit measure.").

[64] *See, e.g.*, David M. Amodio et al., *Neural Signals for the Detection of Unintentional Race Bias*, 15 PSYCHOL. SCI. 88 (2004); Eric J. Vanman et al., *Racial Discrimination by Low-Prejudiced Whites: Facial Movements as Implicit Measures of Attitudes Related to Behavior*, 15 PSYCHOL. SCI. 711 (2004). Whereas the first wave used unobtrusive measures "to assess attitudes, beliefs, and values of which people are aware, but that they may be unwilling to reveal to the investigator," the second wave used "implicit methods . . . to assess attitudes, beliefs, and values of which people are unaware." John F. Kihlstrom, *Implicit Methods in Social Psychology*, *in* THE SAGE HANDBOOK OF METHODS IN SOCIAL PSYCHOLOGY 195, 196 (Carol Sansone et al. eds., 2004). "Response latency measures, which yield evaluations that are unlikely to be controlled, have been heralded because they override the obvious problem of distortion." Laurie A. Rudman, *Sources of Implicit Attitudes*, 13 CURRENT DIRECTIONS PSYCHOL. SCI. 79, 79 (2004).

Implicit measurement techniques supposedly reveal information about mental representations of race. In contrast, neuropsychological techniques, such as brain scans, seek to reveal the physiological bases of psychological and social reactions to racial stimuli. *See* Jennifer L. Eberhardt, *Imaging Race*, 60 AM. PSYCHOL. 181 (2005).

[65] "Most of the [implicit measurement] research that has been conducted has concerned either various forms of priming or the [Implicit Assocation Test]." Fazio & Olson, *supra* note 5, at 300.

[66] *See id.* at 298 ("Probably the most well-known implicit measurement technique is the Implicit Association Test . . . .").

[67] In studies examining implicit stereotypes, researchers study natural associations between social groups and stereotypic traits, and, in studies examining implicit prejudices, researchers study natural associations between social groups and evaluative terms or affect-laden concepts.

subjects are to categorize as good or bad as quickly as possible.[68] A subject's response latency in performing the categorization task is taken as a measure of strength of association between the prime and the target, with faster responses indicating stronger association. For instance, in the context of a racism study, a subject might be primed with pictures of faces of different races followed by positive and negative target adjectives that must be categorized as either "good" or "bad." Slower reaction times in labeling positive adjectives following an African-American prime is taken as evidence of an implicit negative attitude toward African-Americans; so too is faster reaction time in labeling negative traits.[69]

In the Implicit Association Test, or IAT, subjects must pair target concepts with evaluatively charged concepts and then categorize stimuli as falling into one of the competing categories as quickly as possible. As with the priming approach, response latencies serve as the measure of strength of association between these paired concepts.[70] The guiding idea is that we should be able to respond more quickly for items that we associate naturally (e.g., civil procedure + boring, White + pleasant) than those we do not (e.g., civil procedure + exciting, Black + pleasant).[71] For example, studies of

---

[68] *See, e.g.*, Fazio et al., *supra* note 63, at 1013 (describing the basic priming procedure).

[69] Or as Fazio and colleagues explain:

> If the face (prime) automatically activated evaluations from memory, responding in the adjective connotation task should be affected differently for positive versus negative adjectives. Inferences regarding the participant's attitude toward the individuals represented in the photographs can be drawn from the pattern of facilitation. Relatively greater facilitation on negative than positive adjectives when those adjectives are preceded by Black faces than when they are preceded by White faces would be indicative of a more negative attitude toward Blacks.

Id. at 1015.

[70] *See* Anthony G. Greenwald et al., *supra* note 5, at 1464–66 (describing the basic implicit association technique); *see also* Anthony G. Greenwald et al., *supra* note 45, at 8 ("The usefulness of the IAT in measuring association strength depends on the assumption that when the two concepts that share a response are strongly associated, the sorting task is considerably easier than when the two response-sharing concepts are either weakly associated or bipolar-opposed."). Error rates are also taken into account in judging performance on the IAT. For detailed instructions on use of the IAT, see Anthony G. Greenwald et al., *Understanding and Using the Implicit Association Test: I. An Improved Scoring Algorithm*, 85 J. PERSONALITY & SOC. PSYCHOL. 197 (2003). Demonstration versions of the IAT can be found online at https://implicit.harvard.edu/implicit/ (last visited Sept. 28, 2006).

[71] *See* C. Miguel Brendl et al., *How Do Indirect Measures of Evaluation Work? Evaluating the Inference of Prejudice in the Implicit Association Test*, 81 J. PERSONALITY & SOC. PSYCHOL. 760, 761 (2001) ("The IAT is an ingenious use of response competition

implicit racism use some form of the race IAT: subjects first view a set of stimuli with racial connotations (e.g., faces or names associated with different races) that they must categorize as "White" or "Black" by pressing the left or right key on their keyboard as quickly as possible and then view a set of words with positive or negative connotations (e.g., sickness, death, freedom, paradise) that they must categorize as "pleasant" or "unpleasant." Subjects also participate in trials in which they must pair racial and attitudinal stimuli into composite categories. In one trial, the stimuli must be sorted between White + pleasant and Black + unpleasant, and in another trial the stimuli must be sorted between reversed pairings. If subjects take longer to sort stimuli into the reversed pairings of target and evaluative concepts (White + unpleasant, Black + pleasant), this greater response latency is taken to be evidence of pre-existing implicit associations.[72]

Because the IAT and affective-priming studies test associations between social groups and terms having positive and negative evaluative content, researchers contend that these implicit measurements reveal underlying implicit prejudices and not simply implicit stereotypic associations.[73] Thus,

---

designed to measure attitudes indirectly." (citation omitted)); Marco Perugini, *Predictive Models of Implicit and Explicit Attitudes*, 44 BRIT. J. SOC. PSYCHOL. 29, 30 (2005) ("[The IAT] relies on the assumption that, if a target concept and an attribute dimension are highly associated (congruent), the task will be easier, and, therefore, quicker when they share the same response key than when they require a different response key.").

[72] Or as Greenwald and colleagues explain:

> Black + pleasant should be easier than White + pleasant if there is a stronger association between Black Americans and pleasant meaning than between White Americans and pleasant meaning. If the preexisting associations are opposite in direction—which might be expected for White subjects raised in a culture imbued with pervasive residues of a history of anti-Black discrimination—the subject should find White + pleasant to be easier.

Greenwald et al., *supra* note 5, at 1465; *see also* Perugini, *supra* note 71, at 30 ("An IAT score is computed as a function of the difference of the mean response times between the two versions of the combined task. Thus, for instance, respondents will generally be quicker to associate flowers with pleasant, compared to flowers with unpleasant (or, conversely, will be slower to associate insects with pleasant, compared to insects with unpleasant), therefore, revealing a positive implicit attitude towards flowers relative to insects . . . ." (citation omitted)).

[73] The studies can be fashioned, however, to test solely for stereotypic associations, by substituting stereotypic traits for the evaluative concepts. *See* Laurie A. Rudman et al., *Measuring the Automatic Components of Prejudice: Flexibility and Generality of the Implicit Association Test*, 17 SOC. COGNITION 437, 439–40 (1999) ("In prejudice or stereotyping research, [IAT] subjects categorize ingroup and outgroup tokens (*target concepts*), along with stimuli representing the poles of an *attribute dimension*. To assess implicit *prejudice*, the attribute dimension is evaluative and consists of pleasant versus unpleasant words . . . . To assess implicit *stereotypes*, the attribute dimension consists of stereotypic and nonstereotypic words." (citations omitted)); *id.* at 440 ("By varying the

because both the IAT and priming studies employ extremely sensitive measures of response speed, a differential in response times of just milliseconds can be taken as proof of an implicit prejudice toward a particular social group.[74]

Four primary findings from the implicit prejudice research program have energized antidiscrimination law scholars: according to this research: (1) implicit prejudice is pervasive; (2) implicit prejudice is distinct from explicit prejudice; (3) implicit prejudice typically operates beyond conscious control; and (4) implicit prejudice produces discriminatory behavior.[75]

First, in a short period, a tremendous amount of data has been collected showing that people generally respond more quickly when pairing referents to relatively advantaged groups within the U.S. (White, wealthy, healthy, young, heterosexual, Christian, American) with positive adjective terms and referents to relatively disadvantaged groups (non-White, poor, overweight, aged, homosexual, non-Christian, foreign) with negative adjective terms.[76] This pattern holds across social groups: not only do members of dominant groups show alleged in-group preferences, but members of disadvantaged

---

ingroup and outgroup identities, the IAT is easily configured to assess a wide variety of implicit attitudes and stereotypes.").

[74] *See* Chugh, *supra* note 49, at 208 ("[I]t is worth noting that the difference between an implicit pro-white bias and an implicit pro-black bias can be as little as 50 ms.").

[75] *See, e.g.*, Bagenstos, *supra* note 1, at 7–9 (discussing findings of automatic implicit prejudice toward several minority groups in even self-identified egalitarians and experimental research showing linkages between implicit prejudice and behavior); Kang, *supra* note 16, at 1512–14 (discussing finding of pervasive bias operating automatically, dissociation of explicit and implicit bias, and "persuasive evidence that implicit bias against a social category . . . predicts disparate behavior toward individuals mapped to that category"); Kang & Banaji, *supra* note 3, at 1071 ("Most fundamental is the pervasive, replicable, and sometimes large effects of implicit bias . . . . Implicit bias has consequences in the daily activities of our lives.").

[76] *See, e.g.*, William A. Cunningham et al., *Implicit and Explicit Ethnocentrism: Revisiting the Ideologies of Prejudice*, 30 PERSONALITY & SOC. PSYCHOL. BULL. 1332, 1334 (2004) ("[I]mplicit preferences have been found for White over Black Americans, straight over gay, Christian over Jewish, young over old, and American over Soviet" (citations omitted)); Wade C. Rowatt et al., *Patterns and Personality Correlates of Implicit and Explicit Attitudes Toward Christians and Muslims*, 44 J. FOR SCI. STUDY OF RELIGION 29, 39 (2005) ("Christians' implicit and explicit evaluations of the in-group (i.e., Christian) are more favorable than their implicit and explicit evaluations of the out-group (i.e., Muslim)."); Laurie A. Rudman, *Social Justice in Our Minds, Homes, and Society: The Nature, Causes, and Consequences of Implicit Bias*, 17 SOC. JUST. RES. 129, 130 (2004) (noting, in addition to race and age prejudice, evidence of "a similar pattern for prejudices based on religion, physical appearance, and socioeconomic class, as well as sexual orientation" (citations omitted)). Greenwald and Krieger summarize data from a wide variety of IATs. *See* Anthony G. Greenwald & Linda Hamilton Krieger, *Implicit Bias: Scientific Foundations*, 94 CAL. L. REV. 945, 957–58 (2006).

groups allegedly prefer dominant out-groups.[77] This widespread pattern of stronger positive associations with advantaged groups and stronger negative associations with disadvantaged groups seems to signal alarming levels of implicit prejudice within the American public,[78] in stark contrast to the decreasing levels of explicit prejudice found in public opinion surveys.[79]

It is worth noting, however, that other measures fail to find such overwhelming implicit prejudice. For instance, the affective-priming approach "reveals negativity [toward Blacks] in 50 to 60% of White college students, but prejudiced IAT scores are found in 70 to 90% of Whites."[80]

Also, noticeably absent from the list of implicit prejudices revealed by IAT research is implicit male prejudice against women. Although "women strongly prefer female gender when response latency techniques are used, men typically show neutral gender attitudes."[81] Both men and women show a pro-in-group bias in their stereotypic associations (e.g., both men and women more strongly associate their in-group with competence as a trait), but for men these stereotypic associations do not translate into negative implicit

---

[77] Dasgupta, *supra* note 36, at 149 (noting that "a number of studies reveal outgroup favoritism (or sometimes, less ingroup favoritism) in the case of disadvantaged groups, especially when people's attitudes and beliefs are assessed using indirect measures rather than self-report measures").

[78] "Recognition of the pervasiveness of implicit bias lends support to a structural approach to antidiscrimination law." Bagenstos, *supra* note 1, at 10.

[79] Greenwald & Krieger, *supra* note 76, at 955 ("[T]he IAT measures consistently revealed greater bias in favor of the relatively advantaged group (averaging almost three-quarters of respondents across all the topics) than did the explicit measures (for which an average of slightly over one-third of respondents showed bias favoring advantaged groups)."); Rudman, *supra* note 76, at 130 ("[I]f researchers were to rely solely on self-report measures, they would be tempted to conclude that prejudice has become, if not outdated, at least unfashionable .... However, this does not mean that the problem of bigotry has been solved, for when attitudes are measured using methods that do not rely on respondents' willingness or ability to report their opinions, the persistence of prejudice and stereotypes is routinely exposed." (citations omitted)).

[80] Michael A. Olson & Russell H. Fazio, *Relations Between Implicit Measures of Prejudice: What Are We Measuring?*, 14 PSYCHOL. SCI. 636, 636 (2003) (citations omitted).

[81] Laurie A. Rudman & Stephanie A. Goodwin, *Gender Differences in Automatic In-Group Bias: Why Do Women Like Women More Than Men Like Men?*, 87 J. PERSONALITY & SOC. PSYCHOL. 494, 494–95 (2004) (citations omitted). That is, women are "implicitly sexist," but men are not. *Id.* at 495. However, Rudman and Kilianski did find evidence in both men and women of implicit prejudice toward female authorities. *See* Laurie A. Rudman & Stephen E. Kilianski, *Implicit and Explicit Attitudes Toward Female Authority*, 26 PERSONALITY & SOC. PSYCHOL. BULL. 1315, 1324–26 (2000).

attitudes toward women.[82] "The fact that women show stronger automatic in-group bias than men is provocative, because it suggests a reversal of sexism at the implicit level."[83]

Second, an individual's scores on implicit and explicit measures of prejudice often diverge, suggesting that many people who consciously define and express themselves as egalitarians may nevertheless be implicit bigots.[84] It is this disjunction between conscious and unconscious attitudes that often elicits the most discomfort and disbelief among subjects, causing subjects to question the validity of the IAT.[85] And it is also this disjunction that makes pervasive implicit prejudice so alarming from a legal perspective, because implicit prejudice is not just hidden from public view—it is hidden from our own personal introspection as well. If implicit prejudice operates independently of explicit prejudice, then we can no longer use direct or circumstantial evidence of intentional conduct as a good indicator of prejudice, but we must instead delve into the inner workings of the mind using the IAT or some other means of implicit measurement.[86]

---

[82] *See, e.g.*, Jennifer A. Richeson & Nalini Ambady, *Who's in Charge? Effects of Situational Roles on Automatic Gender Bias*, 44 SEX ROLES 493 (2001). Rudman and Goodwin unconfounded gender attitudes from stereotypic beliefs in a series of IAT studies and found that men's "in-group bias is surprisingly frail and that women's in-group bias is particularly strong at the implicit level." Rudman & Goodwin, *supra* note 81, at 506. "[O]ur experiments suggest that in the absence of specific power manipulations, women strongly implicitly prefer their own gender, whereas men do not." *Id.*

[83] Rudman & Goodwin, *supra* note 81, at 508.

[84] *See* Greenwald et al., *supra* note 5, at 1477 ("It is clear that these implicit-explicit correlations should be taken not as evidence for convergence among different methods of measuring attitudes but as evidence for divergence of the constructs represented by implicit versus explicit attitude measures."); Rudman et al., *supra* note 73, at 461 ("Consistent with prior research, the IAT and self-report measures . . . were weakly or unreliably related to one another, indicating that the two types of methods assess independent constructs." (citations omitted)).

[85] *See* Nilanjana Dasgupta et al., *The First Ontological Challenge to the IAT: Attitude or Mere Familiarity?*, 14 PSYCHOL. INQUIRY 238, 239 (2003) (describing subject reactions to the racial IAT). This surprising finding may also partially explain the popularity of the IAT method among researchers. *See* Allen R. McConnell & Jill M. Leibold, *Relations Among the Implicit Association Test, Discriminatory Behavior, and Explicit Measures of Racial Prejudice*, 37 J. EXPERIMENTAL SOC. PSYCHOL. 435, 436 (2001) ("Social psychologists who study group prejudice have been drawn to the IAT because of its large effect size and because even people who know that the IAT assesses group prejudice still reliably produce the IAT effect, indicating its robustness and apparent imperviousness." (citations omitted)).

[86] *See* Bagenstos, *supra* note 1, at 10–11 (discussing the proof problems presented by implicit bias and prejudice); *see also* Jason A. Nier, *How Dissociated Are Implicit and*

Unfortunately, for purposes of clarity, we shall soon discover that the degree to which implicit and explicit attitudes overlap remains a point of considerable uncertainty.[87]

Third, and related to the second point, "[i]mplicit attitudes ... are disengaged from conscious thought and are unlikely to shift in response to the willful call for change."[88] Nevertheless, even minor alterations in the implicit association testing situation can lead to significant changes in IAT response patterns,[89] and implicit attitudes have been shown to change in response to a variety of interventions.[90] In addition, motivation to control

---

Explicit Racial Attitudes? A Bogus Pipeline Approach, 8 GROUP PROCESSES & INTERGROUP RELATIONS 39, 49 (2005) (suggesting that the dissociation between implicit and explicit measures of social group attitudes is due to shortcomings in explicit measures and that implicit measures more accurately capture these attitudes by avoiding social desirability response problems).

[87] *See* Kurt A. Boniecki & Julia Zuwerink Jacks, *The Elusive Relationship Between Measures of Implicit and Explicit Prejudice*, 26 REPRESENTATIVE RES. SOC. PSYCHOL. 1, 2 (2002) ("The degree to which implicit and explicit prejudice are related has been a matter of some debate."); Timothy D. Wilson & Elizabeth W. Dunn, *Self-Knowledge: Its Limits, Value, and Potential for Improvement*, 55 ANN. REV. PSYCHOL. 493, 502 (2004) ("[M]any studies have found low correlations between explicit and implicit measures of attitudes, though some have found higher degrees of correspondence." (citations omitted)).

[88] Mahzarin R. Banaji, *The Opposite of a Great Truth Is also True: Homage to Koan #7*, *in* PERSPECTIVISM IN SOCIAL PSYCHOLOGY 127, 135 (John T. Jost et al. eds., 2004); *see also* Robyn K. Mallett et al., *What Intergroup Relations Research Can Tell Us About Coalition Building*, 12 WASH. & LEE J. C. R. & SOC. JUST. 5, 10–11 (2005) ("[I]mplicit prejudice is an attitude that we are often unaware that we hold, we cannot consciously examine, and that is largely out of our conscious control." (footnote omitted)). For concerns about the validity of the assumption that conscious knowledge does not influence implicit attitudes as measured by the IAT, see *infra* note 236.

[89] *See* Banaji, *supra* note 88, at 135 ("I was unprepared for data that showed that the influence that minor variations in social situations, such as the presence or absence of a person, can play in defining the attitude object itself—the different construals possible of seemingly the same attitude object."); *see also* Devine, *supra* note 4, at 758 ("[I]t is clear that context manipulations produce replicable patterns of moderation of implicit biases.").

[90] *See* Irene V. Blair et al., *Imagining Stereotypes Away: The Moderation of Implicit Stereotypes Through Mental Imagery*, 81 J. PERSONALITY & SOC. PSYCHOL. 828, 837 (2001) ("Five experiments provided compelling evidence for the moderating influence of mental imagery on implicit stereotypes."); Nilanjana Dasgupta & Anthony G. Greenwald, *On the Malleability of Automatic Attitudes: Combating Automatic Prejudice with Images of Admired and Disliked Individuals*, 81 J. PERSONALITY & SOC. PSYCHOL. 800, 806 (2001) ("Although automatic attitudes have been previously conceptualized as relatively immutable, the present research provides new evidence suggesting that automatic preference and prejudice may indeed be malleable." (citations omitted)); Tiffany A. Ito et al., *The Influence of Facial Feedback on Race Bias*, 17 PSYCHOL. SCI. 256, 259 (2006) ("The present research demonstrates that repeatedly viewing Black faces while being surreptitiously induced to smile diminishes implicit racial bias."); Laurie A. Rudman et

prejudice can lead to correction of automatic attitudes,[91] and social-cognitive goals appear to mediate the types of associations automatically brought to mind.[92] As with the relation between explicit and implicit prejudice, we are not dealing with settled science here: the degree to which implicit attitudes operate automatically and consistently across contexts is a source of controversy among researchers and we return to these issues when we discuss the external validity of implicit-prejudice research.[93]

Fourth, implicit prejudice manifests itself in acts of discrimination, at least under some conditions.[94] Although very few studies have examined the relation between implicit prejudice and "macro-level" behaviors (e.g., discrimination in employment decisions), a greater number of studies have demonstrated a correlation between implicit prejudice and "micro-level" behaviors (e.g., body posture and other forms of nonverbal conduct in the presence of a minority).[95] A recent meta-analysis of IAT studies confirmed

---

al., *"Unlearning" Automatic Biases: The Malleability of Implicit Prejudice and Stereotypes*, 81 J. PERSONALITY & SOC. PSYCHOL. 856, 864 (2001) ("These findings strongly support the hypothesis that people can 'unlearn' both explicit and implicit prejudice in real-world contexts.").

[91] *See* Fazio & Olson, *supra* note 5, at 319 ("In a variety of studies, the more motivated show evidence of having 'corrected' for their automatically activated attitudes."). That this research employs scales ostensibly measuring "motivation to control prejudice" illustrates, again, the alacrity with which social psychologists affix value-laden labels that obscure instead of illuminate the diverse unobservable processes that might be engaged. As Tetlock and Arkes note, these scales may tap into a difficult-to-disentangle mix of competing views on: (a) what constitutes prejudice; (b) what others think constitutes prejudice; and (c) the appropriateness of engaging in conduct that might confirm to oneself or others that one is "prejudiced." *See* Tetlock & Arkes, *supra* note 29, at 315. The root misconception here is that prejudice is a natural science construct on which there is virtual unanimity (e.g., DNA) instead of a social-political construct that takes on different meanings for different people at different points in history.

[92] *See* Mary E. Wheeler & Susan T. Fiske, *Controlling Racial Prejudice: Social-Cognitive Goals Affect Amygdala and Stereotype Activation*, 16 PSYCHOL. SCI. 56, 61 (2005) ("Together the experiments show that a stereotyped or prejudiced response to an out-group member requires, at minimum, that the stimulus (a photo in this case) be processed deeply enough that it represents a social target . . . . Most important, the results show that perceivers can change the social context in which they view a target person and thereby affect out-group perception measurable in both the brain and reaction time behavior.").

[93] *See infra* Part III.D.

[94] *See* Dasgupta, *supra* note 36, at 163 ("[I]t is also clear that people's implicit attitudes and beliefs toward in- and outgroups affect specific types of behaviors, some of which may operate without social actors' awareness or control; but it is also evident that implicit biases do not always result in discriminatory action in an obligatory fashion.").

[95] *See* Jonathan C. Ziegert & Paul J. Hanges, *Employment Discrimination: The Role of Implicit Attitudes, Motivation, and a Climate for Racial Bias*, 90 J. APPLIED PSYCHOL. 553, 553 (2005) ("Although research has documented that these implicit measures

the moderate predictive value of IAT scores with respect to behaviors purportedly diagnostic of prejudice or stereotyping—at least in artificial laboratory settings in which participants believe their judgments to be unaccountable and hypothetical as opposed to judgments with real consequences for which they may be accountable.[96]

C. *Implicit Prejudice as an Emerging Legal Concept*

Legal scholars' embrace of implicit prejudice research constitutes the latest phase in an effort to bring antidiscrimination law's psychological assumptions into line with empirical reality.[97] A primary concern of legal

---

correlate with other attitudes and predict microlevel behavior, there is currently little evidence indicating that such implicit attitudes are useful for predicting more macrolevel behavior, such as discriminatory hiring decisions."); *see also id.* at 561 ("[I]mplicit racism interacted with a climate for racial bias to predict discrimination . . . . This is one of the first studies to demonstrate that the IAT can predict racially biased discriminatory actions.").

[96] *See* Poehlman et al., *supra* note 62, at 21 ("This meta-analysis indicates that IAT measures are significant predictors of criterion measures (average $r = .27$) . . . . Explicit (i.e., self-report) measures were also good predictors of criterion measures, and in fact performed significantly better overall than IAT measures did (average $r = .35$)."); *id.* at 28 ("IAT measures significantly out-predicted explicit measures in the domain of prejudice and stereotyping."); *see also id.* at 62 tbl.3 (reporting mean IAT-criterion correlation for prejudice/stereotyping studies of $r = .25$). For social science studies, Cohen characterizes a correlation coefficient (r) of .30 as moderate ($r = .10$ represents a small effect and $r = .50$ represents a large effect). *See* JACOB COHEN, STATISTICAL POWER ANALYSIS FOR THE BEHAVIORAL SCIENCES 78–81 (2d ed. 1988). With an $r = .30$, an independent variable explains 9% of the variance in a dependent variable; an $r = .25$ explains 6.25% of the variance.

[97] *See* Krieger & Fiske, *supra* note 25, at 1003 ("[T]here has emerged in the past ten years a school of legal scholarship exploring the implications of insights emerging from psychological science for antidiscrimination law and policy" (footnote omitted)); Ann C. McGinley, *¡Viva La Evolución!: Recognizing Unconscious Motive in Title VII*, 9 CORNELL J.L. & PUB. POL'Y 415, 491 (2000) ("If Title VII is to fulfill its purpose, it must define discrimination in accordance with scientific understanding. The law can no longer limit its definition of discrimination to conscious discriminatory behavior; the definition should also include behavior that is rooted in unconscious prejudice."); *see also* Linda Hamilton Krieger, *The Intuitive Psychologist Behind the Bench: Models of Gender Bias in Social Psychology and Employment Discrimination Law*, 60 J. SOC. ISSUES 835, 836 (2004) ("One important reason why working mothers often find it difficult to win meritorious sex discrimination cases is that inaccurate assumptions about the nature of inter-group perception and judgment still permeate the courts.").

The seminal paper in this effort is Charles Lawrence's *The Id, the Ego, and Equal Protection: Reckoning with Unconscious Racism*. *See* Lawrence, *supra* note 17. *See* Bagenstos, *supra* note 1, at 6 ("In the past decade, a number of scholars have taken up Professor Lawrence's project and given added momentum to the notion that unconscious or subtle bias is a major contributor to today's problems of workplace inequality."). For

scholars is that the law's requirement of proof of an intentional mental state as a cause of discrimination, unless broadly construed, fails to capture many instances of discrimination caused by automatic stereotyping processes and in-group preferences; hence, proof of a discriminator's conscious awareness of the causes of discrimination should not be required.[98] Implicit prejudice research heightens concern about the law's emphasis on conscious discrimination because this new research provides evidence of implicit group biases that cannot be detected with direct self-report measures of explicit prejudice, and it suggests that implicit prejudice is widespread and not dependent on conscious animus toward minorities.[99] The notion of implicit prejudice thus does triple duty: (1) it explains how inequalities between advantaged and disadvantaged groups can persist despite the well-documented decline in overt expressions of racism and sexism;[100] (2) it performs this delicate explanatory task without blaming the victim (indeed, it expands the set of victims to include those who do not yet realize they have been victimized by subtle forms of prejudice); and (3) it casts doubt on the

---

citations to several of the works that followed Lawrence's lead, see Marc R. Poirier, *Is Cognitive Bias at Work a Dangerous Condition on Land?*, 7 EMP. RTS. & EMP. POL'Y J. 459, at 461 n.9 (2004). Another particularly influential law review paper in this line of work is Linda Hamilton Krieger's, *The Content of Our Categories: A Cognitive Bias Approach to Discrimination and Equal Employment Opportunity*, 47 STAN. L. REV. 1161 (1995).

[98] *See, e.g.*, Krieger & Fiske, *supra* note 25, at 1004 ("[S]cholars, who include both lawyers and social psychologist . . . generally advocate a causation-based, rather than an intent-based, understanding of the antidiscrimination principle. They also support an expansive application of disparate impact theory in cases involving subjective decision-making systems or other processes or criteria that tend to systematically deprive historically marginalized groups of employment opportunities" (footnotes omitted)); David Benjamin Oppenheimer, *Negligent Discrimination*, 141 U. PA. L. REV. 899, 899 (1993) ("[A] theory of discrimination liability that focuses on intentional wrongdoing will inevitably miss the mark."); Sturm, *supra* note 1, at 478 ("Problems of gender-based exclusion and unfairness will persist without external legal regulation, but the first generation form of regulation is inadequate to address the complexities of second generation bias."); Rebecca Hanner White & Linda Hamilton Krieger, *Whose Intent Matters? Discrimination in Multi-Actor Employment Decision Making*, 61 LA. L. REV. 495, 498 (2001) ("We believe that Title VII *should* be interpreted, and the Supreme Court's decisions can and *should* be read, as rejecting a requirement of conscious intent." (footnote omitted)). The main legal prescription drawn from the implicit prejudice research by the researchers themselves is that the law's intentionality requirement within antidiscrimination law, and perhaps other areas of the law, should be rescinded. *See supra* note 14.

[99] *See generally* Kang, *supra* note 16, at 1506–14.

[100] *See, e.g.*, Dovidio, *supra* note 42, at 845 ("[A]lthough overt expressions of prejudice have declined steadily and significantly over time, subtle—often unconscious and unintentional—forms continue to exist.").

ability of current antidiscrimination law to further mitigate inequality and highlights a reformist agenda.

It is difficult to overstate the legal significance of this new research if it correctly diagnoses the pervasiveness and potency of implicit prejudice and related discriminatory tendencies. As Professor Kang recently put it, "[a] vast intellectual agenda opens when we start probing what this new knowledge might mean for law."[101] Measures of implicit prejudice might be used to test legislators for racial bias,[102] to screen possible employees,[103] and to root out prejudice among existing employees.[104] The threat of implicit prejudice bolsters the argument for abolition of the peremptory challenge[105] and for recognition of a disparate-impact claim under Title VI for discriminatory provision of medical services.[106] Implicit prejudice even provides a rationale for greater FCC regulation of the media, which broadcasts many crime stories involving minorities that may strengthen implicit negative attitudes toward minority groups.[107]

Already, implicit prejudice research has been put to practical legal use. Eisenberg and Johnson, to raise awareness of racial bias in death penalty cases, used measures of implicit bias to demonstrate that even capital defense lawyers harbor negative implicit attitudes toward minorities,[108] and experts

---

[101] Kang, *supra* note 16, at 1494.

[102] *See* Saujani, *supra* note 13, at 420.

[103] *See* IAN AYRES, PERVASIVE PREJUDICE? 424 (2001) ("Implicit attitude testing might also itself be used as a criterion for hiring both governmental and nongovernmental actors.").

[104] *See* Deana A. Pollard, *Unconscious Bias and Self-Critical Analysis: The Case for a Qualified Evidentiary Equal Employment Opportunity Privilege*, 74 WASH. L. REV. 913, 915–16 (1999).

[105] *See* Page, *supra* note 12, at 245–46; *see also* John J. Francis, *Peremptory Challenges, Grutter, and Critical Mass: A Means of Reclaiming the Promise of* Batson, 29 VT. L. REV. 297, 298 (2005) (using implicit prejudice work to support a proposal to exclude minority criminal defendants from the *Batson* rule). Page recognizes that abolition of the peremptory challenge in the near future is unlikely, and so he also calls for changes in the application of *Batson* to peremptory challenges. *See* Page, *supra* note 12, at 245–61.

[106] *See* Michael S. Shin, Comment, *Redressing Wounds: Finding a Legal Framework to Remedy Racial Disparities in Medical Care*, 90 CAL. L. REV. 2047, 2081–82 (2002).

[107] *See* Kang, *supra* note 16, at 1549–53.

[108] Theodore Eisenberg & Sheri Lynn Johnson, *Implicit Racial Attitudes of Death Penalty Lawyers*, 53 DEPAUL L. REV. 1539, 1553 (2004); *see also* Milton D. Jones, *The First Two Seconds: Racial and Gender Bias in Bankruptcy Administration*, 2006 No. 02 NORTON BANKR. L. ADVISER 3 (arguing, from the IAT research, that implicit biases adversely affect trustee treatment of female and minority debtors). Given that the

have begun incorporating implicit-prejudice research into their testimony to create a social framework that supports plaintiffs' claims of subtle discrimination.[109] Most notably, the sociologist William Bielby, who often testifies on behalf of plaintiffs in employment discrimination class actions,[110] has filed several expert reports with courts in which he opines that the great majority of White male employees at the defendant organizations suffer from implicit biases that likely led to discrimination against the plaintiff classes.[111]

Most of these legal uses rest on disturbingly uncritical assessments of the psychological arguments advanced by proponents of the IAT. We shall show in the next sections that it would be a big mistake for the legal community to accept these claims at face value, so big a mistake that it is necessary to delve into the scientific controversies surrounding implicit prejudice in greater detail than has thus far been the case in law-review literature. Much hinges on how one reads the technical literature that purports to gauge the psychological processes underlying IAT scores, the robustness of the prejudicial attitudes supposedly revealed by the IAT, and the reliability of the connection between these prejudices and discriminatory behavior. Close

---

inventors of tests of implicit prejudice often fail their own tests and resort to efforts to debias themselves, we should expect even those with the purest of conscious intentions often to fall short. In an article for *The Harvard University Gazette*, William J. Cromie states:

> There's no way to wipe out all the years of evolution during which humans and their ancestors learned to fear the unfamiliar, to be ready to run or fight at any threat. But brains are flexible enough to be altered by experience. 'If this were not so, we would not have seen the reduction of bias in students who worked with a black researcher,' Banaji points out. We shouldn't expect such changes to last long, but it does cause Banaji to feel optimistic about making behavioral changes that, though small and temporary, are real.

William J. Cromie, *Brain Shows Unconscious Prejudices: Fear Center Is Activated*, HARV. UNIV. GAZETTE, July 17, 2003, *available at* http://www.news.harvard.edu/gazette/archives.html (follow "2003" hyperlink; then follow "July 17" hyperlink; then follow "Science/Research" hyperlink).

[109] *See* William T. Bielby, *Can I Get a Witness? Challenges of Using Expert Testimony on Cognitive Bias in Employment Discrimination Litigation*, 7 EMP. RTS. & EMP. POL'Y J. 377, 390 (2003).

[110] Michael Orey, *White Men Can't Help It*, BUS. WK., May 15, 2006, at 54, *available at* http://www.businessweek.com/magazine/content/06_20/b3984081.htm?campaign_ id=search ("Sociologist William T. Bielby is the leading courtroom proponent of a simple but powerful theory: 'unconscious bias.' He contends that White men will inevitably slight women and minorities because they just can't help themselves. So he tries to convince judges that no evidence of overt discrimination—no smoking gun memo, for instance—is needed to prove a case.").

[111] *See id.* (noting that Professor Bielby has been an expert witness "in dozens of major cases, including those currently pending against Wal-Mart, FedEx, Johnson & Johnson, and Cargill").

inspection will yield an anti-climactic conclusion: if the goal is application to the law, implicit prejudice research does not yet pass minimum standards of reliable science.

III. EXAMINING THE SCIENTIFIC VALIDITY OF IMPLICIT PREJUDICE RESEARCH

Social scientific research has traditionally been evaluated against four benchmarks of validity: (1) construct validity, which asks whether the unobservable target of research interest, such as prejudice, has been properly operationalized or translated into an observable variable that can be either manipulated or measured for purposes of hypothesis-testing; (2) internal validity, which asks whether an observed relationship between variables represents a causal relationship; (3) statistical conclusion validity, which asks whether observed covariation between variables is real or spurious; and (4) external validity, which asks whether results obtained in one setting generalize to other populations, situations, and times.[112] Invalidity of the first type renders research results incomplete, unconvincing or, at worst, useless, because the variables studied do not adequately match the phenomena of theoretical or practical interest. Invalidity of the second or third type makes it impossible to draw valid conclusions about causal or correlational linkages. Invalidity of the fourth type restricts the domains to which valid inductive inferences may be drawn from the research.[113]

Basic research, in which the goal is to test competing theories, emphasizes construct, internal, and statistical conclusion validity.[114] Applied

---

[112] *See* THOMAS D. COOK & DONALD T. CAMPBELL, QUASI-EXPERIMENTATION: DESIGN & ANALYSIS ISSUES FOR FIELD SETTINGS 37–39 (1979).

[113] Scientific validity is often a matter of degree, particularly for construct and external validity, and it is difficult to identify, a priori, the point at which validity problems or questions become so great that a study or body of work is deemed scientifically invalid.

[114] *See* COOK & CAMPBELL, *supra* note 112, at 83 ("The priority among validity types varies with the kind of research being conducted . . . . For investigators with theoretical interests our estimate is that the types of validity, in order of importance, are probably internal, construct, statistical conclusion, and external validity."). Some methodologists distinguish between internal validity, which can encompass statistical conclusion validity, and external validity, which can encompass construct validity. *See id.* at 80–82. Campbell and Stanley offer the classic statement on internal and external validity and their relation:

> *Internal validity* is the basic minimum without which any experiment is uninterpretable: Did in fact the experimental treatments make a difference in this specific experimental instance? *External validity* asks the question of *generalizability*: To what populations, settings, treatment variables, and

research, in which the goal is to inform decision-makers about existing problems, places much greater importance on external validity.[115] Applied legal research must also contend with the validity requirements imposed by the Supreme Court in the *Daubert* trilogy: to be admissible in litigation, scientific and other forms of expert evidence must be reliable and relevant and must "fit" the facts of the case.[116]

These legal requirements for evidentiary uses of social science research largely incorporate the four categories of scientific validity, but the fit requirement places a heavier weight on construct validity and external validity.[117] Even when psychological research on phenomena labeled bias, prejudice and discrimination is internally and statistically valid, if the phenomena so labeled do not correspond to phenomena that judges consider bias, prejudice, and discrimination, then the psychological research lacks construct validity as a legal matter and cannot fit the case at hand. Likewise, if psychologists can only demonstrate implicit prejudice and discrimination in contrived settings that eliminate complicating factors likely to exist in the

---

> measurement variables can this effect be generalized? Both types of criteria are obviously important, even though they are frequently at odds in that features increasing one may jeopardize the other. While *internal validity* is the *sine qua non*, and while the question of *external validity*, like the question of inductive inference, is never completely answerable, the selection of designs strong in both types of validity is obviously our ideal.

DONALD T. CAMPBELL & JULIAN C. STANLEY, EXPERIMENTAL AND QUASI-EXPERIMENTAL DESIGNS FOR RESEARCH 5 (1963).

[115] *See* COOK & CAMPBELL, *supra* note 112, at 83 ("The priority ordering for many applied researchers is something like internal validity, external validity, construct validity of the effect, statistical conclusion validity, and construct validity of the cause."); *see also* Krieger & Fiske, *supra* note 25, at 44 ("Psychological science is basic research, often conducted in the laboratory. Meanwhile, legal disputes concern events in the real world. As such, questions of 'external' or 'field' validity inevitably arise." (footnote omitted)).

[116] *See* Kumho Tire Co. v. Carmichael, 526 U.S. 137, 141 (1999); Gen. Elec. Co. v. Joiner, 522 U.S. 136, 142 (1997); Daubert v. Merrell Dow Pharm., Inc., 509 U.S. 579, 589–91 (1993). For a discussion of the impact of *Daubert* and its progeny on psychological evidence, see generally David L. Faigman & John Monahan, *Psychological Evidence at the Dawn of the Law's Scientific Age*, 56 ANN. REV. PSYCHOL. 631 (2004).

[117] The status of empirical studies vis-à-vis the specific criteria suggested in *Daubert* as indicators of scientific validity—testability, error rate, peer review status, and general acceptance—depend on the status of the empirical studies vis-à-vis the four categories of scientific validity described in the text. *See generally* DAVID L. FAIGMAN ET AL., SCIENCE IN THE LAW: STANDARDS, STATISTICS AND RESEARCH ISSUES 115–49 (2002). Basic research that suffers from flaws of construct, internal, or statistical conclusion validity is unlikely to satisfy some or all of the *Daubert* factors, and applied research that suffers from flaws of construct or external validity is unlikely to "fit" the case at hand.

applied domains of interest, then this psychological research will again fail to fit the case at hand, this time due to lack of external validity.[118]

In the following sections, we demonstrate that implicit prejudice research suffers from serious shortcomings within each category of validity. Unresolved questions exist about how psychologists have chosen to define and measure implicit prejudice and about how well the results of psychologists' mindreading efforts generalize beyond the laboratory. Further, even if we view the implicit prejudice research exclusively as a program of basic research with no applied aspirations, we find that the claims for widespread implicit prejudice and concomitant discriminatory behavior being drawn from the IAT research are much shakier than often reported in the law reviews and in the media. No scientific consensus yet exists with respect to the causes or consequences of the "IAT effect" (that is, more time needed by subjects taking the IAT to respond to pairings of minorities with positive attributes), and the implicit prejudice explanation for this effect is merely the least flattering to subjects, and most useful to reformers, of several competing explanations.[119]

---

[118] External validity constraints greatly diminish the ability of an expert to draw general causation conclusions, much less specific causation conclusions, from social science research in order to develop a social framework or social fact. *See* Faigman & Monahan, *supra* note 116, at 652 ("If courts treat social frameworks such as the 'battered woman syndrome' or the 'rape trauma syndrome' the way they treat what might be called medical frameworks, they should require that both general causation and specific causation be demonstrated independently under *Daubert*. Moreover, without adequate proof of general causation, no testimony at all on specific causation will be permitted. Even with proof of general causation, proof must still be forthcoming that specific causation is within the capabilities of the science and the proffered expert. Much of psychology is likely to struggle to meet these standards, as medicine has struggled as well." (citation omitted)).

[119] William J. Gehring et al., *Thinking About Interracial Interactions*, 6 NATURE NEUROSCIENCE 1241, 1242 (2003) ("[T]here are several alternative explanations for IAT effects that do not involve unconscious evaluations or prejudice." (footnote omitted)). Indeed, even leading implicit prejudice researchers have conceded the point as recently as 2001: "Research on alternative theoretical interpretations has not yet progressed enough to establish any theoretical interpretation of the IAT effect." Anthony Greenwald & Brian Nosek, *Health of the Implicit Association Test at Age 3*, 46 ZEITSCHRIFT FUR EXPERIMENTELLE PSYCHOLOGIE, 85, 90 (2001). We show here that the theoretical picture has only grown murkier since then—raising serious questions about the legal standing of this entire line of work.

A. *Construct Validity and Metric Meaning: Do Tests of Implicit Prejudice Measure What They Purport to Measure?*

Psychological constructs have always had a shaky ontological status; not directly observable, they must be inferred from behavior.[120] For instance, we infer preference from choice, intelligence from ability tests, and, for the IAT, implicit prejudice from reaction times on word association tests. To close the gap between psychological construct and observed behavior, psychologists must provide a theory about how the observed behavior reflects an underlying mental state or process, and this theory, like any theory, must be subjected to testing and debate to establish the validity of the psychology-behavior connection.[121] Given the uncertainties inherent in such gap-filling, it is not unusual for controversies to stretch over many decades about what particular psychological measures actually gauge, be they IQ, polygraph, or Rorschach inkblot tests. The still youthful IAT has already begun to deliver on the promise of its distinctive measurement controversies.

Dating back to a classic paper by Cronbach and Meehl published in 1955, the standard scientific answer to this conundrum is to invoke the logic of construct validation.[122] In a nutshell, psychologists are justified in concluding that a test taps into an unobservable construct to the degree that construct, and only that construct, can explain the patterns of intercorrelations between the test and other observable indicators. In more

---

[120] Michell summarizes the basic problem:

> The attributes that psychometricians aspire to measure are not directly observable (i.e. claims made about them can only [at present] be tested by first observing something else and making inferences). What psychometricians observe are the responses made to test items. Intellectual abilities, personality traits and social attitudes are theoretical attributes and test scores are taken to be *quantitative* relationships (i.e. functional relationships between quantitative attributes).

Joel Michell, *Normal Science, Pathological Science and Psychometrics*, 10 THEORY & PSYCHOL. 639, 648 (2000).

[121] *See* Drew Westen & Robert Rosenthal, *Quantifying Construct Validity: Two Simple Measures*, 84 J. PERSONALITY & SOC. PSYCHOL. 608, 609 (2003) ("[C]onstruct validation is always theory dependent. A statement about the validity of an instrument is a statement about the extent to which its observed associations with measures of other variables match theoretical predictions about how it should be associated with those variables." (citation omitted)). Psychology is not alone, of course, in having to close gaps between theoretical constructs and empirically observable states or events. The hardest of the hard sciences, physics, employs numerous theoretical constructs and laws that cannot be directly observed but must be inferred from observable facts or regularities.

[122] Lee J. Cronbach & Paul E. Meehl, *Construct Validity in Psychological Tests*, 52 PSYCHOL. BULL. 281, 281 (1955).

technical language, psychological constructs are defined by their location in a constellation (nomological net) of related constructs. The construct's theoretical location should lead to predictions about the convergence and divergence of the measure of the target construct with those of other constructs: "Researchers typically establish construct validity by presenting correlations between a measure of a construct and a number of other measures that should, theoretically, be associated with it (convergent validity) or vary independently of it (discriminant validity)."[123]

1. *Do Implicit Measures of Prejudice Converge as Theory Says They Should?*

The least controversial method of construct validating implicit measures of prejudice should be against other implicit measures of prejudice. As a matter of simple logic, if an implicit measure of prejudice is tapping into a previously hidden reservoir of hostility toward a minority group, this measure should both correlate with itself over time (test-retest) as well as with related measures of the same target construct.[124] The implicit prejudice research program struggles, however, to pass even these minimalist tests. Test-retest reliability coefficients for the IAT are at the low end of the respectable range for measures of ostensibly stable individual-difference constructs, but even this anemic claim cannot be sustained for other implicit measures.[125] And the two most popular methods for measuring implicit

---

[123] Westen & Rosenthal, *supra* note 121, at 608. Some use the term "convergent validity" in the narrower sense of "associations among alternative measures of the same (rather than related) constructs." Duane T. Wegener & Leandre R. Fabrigar, *Constructing and Evaluating Quantitative Measures for Social Psychological Research: Conceptual Challenges and Methodological Solutions*, *in* THE SAGE HANDBOOK OF METHODS IN SOCIAL PSYCHOLOGY, *supra* note 64, at 145, 161.

[124] When researchers claim special access to hidden prejudicial states, and claim that these hidden states have true causal force, while explicit, non-racist explanations for seemingly discriminatory behavior can be dismissed as nothing more than illusions of free will or symptoms of a false consciousness, then the best avenue for falsifying researchers' claims about the existence and power of these hidden psychological states is to examine alternative measures of these hidden states for convergence. If advocates of the IAT reject convergence with other existing measures of implicit prejudice as an essential test of construct validity, then it is unclear how exactly the IAT could ever be convincingly validated as a measure of the implicit prejudice construct.

[125] Fazio & Olson, *supra* note 5, at 311 ("Test-retest reliability for the IAT does tend to reach a respectable level of .6 or higher, as also is true for the name-letter preference task. However, the few reports regarding test-retest reliability for various priming measures have ranged from abysmally low to moderate levels of ~.5." (citations omitted)); Melanie C. Steffens & Axel Buchner, *Implicit Association Test: Separating*

prejudice—the IAT and affective priming—do not correlate nearly as highly as one would expect if the measures were tapping into the same stable, unitary psychological construct.[126] Are we to believe that there are many hidden pockets of unconscious prejudice within the mind, unrelated to one another, and differentially accessed by the IAT and affective priming (notwithstanding that proponents of the tests claim to tap into similar mental processes)?[127] Should we accept the measurement error defense that all is well and that, after we make generous statistical corrections for unreliability, the modest correlations start looking respectable?[128] Or should we conclude that psychological processes unrelated to unconscious prejudice are shaping responses to these tests?[129] This list of unknowns is an early warning sign of the scientific immaturity of the implicit prejudice research program.

2. *Do Implicit and Explicit Measures of Prejudice Co-Vary in Theoretically Predictable Ways—And Where Is the Theory Anyway?*

A second seemingly straightforward strategy of construct validation is to examine the degree to which self-report scales designed to measure explicit prejudice correlate with measures of implicit prejudice. But appearances can

---

Transsituationally Stable and Variable Components of Attitudes Toward Gay Men, 50 EXPERIMENTAL PSYCHOL. 33, 45 (2003) ("[T]he main finding of our experiments is that the test-retest correlation of an IAT assessing implicit attitudes, as obtained directly or estimated in a structural equation model, was rather low ($.50 < r < .62$), even when the IAT was replicated immediately."); *id.* at 34 tbl.1 (summarizing test-retest correlations across several studies). Most psychometricians agree that the test-retest reliability (the power of a test to predict itself) sets a plausible upper bound on the power of the test to predict other criterion variables.

[126] Fazio & Olson, *supra* note 5, at 311 ("One of the most disturbing trends to emerge in the literature on implicit measures is the many reports of disappointingly low correlations among the measures."). *But see* William A. Cunningham et al., *Implicit Attitude Measures: Consistency, Stability, and Convergent Validity*, 12 PSYCHOL. SCI. 163, 167 (2001) (reporting more robust correlations using a latent variable approach that supposedly controls for substantial measurement error in implicit measures).

[127] Crandall and Eshleman distinguish a variety of prejudices that express themselves to varying degrees depending on the balance of social forces promoting tolerance or bigotry. *See* Crandall & Eshleman, *supra* note 43, at 437. Presumably, however, implicit measures aimed at tapping the same prejudice in the same context should converge even under this model, *ceteris paribus*.

[128] For the most aggressive affirmation of this defense, see Cunningham et al., *supra* note 126, at 167. It is instructive, though, that IAT advocates overlook the complications created by measurement error when they treat the zero-point on their scale (no differential reaction times to Whites and Blacks) as the basis for drawing conclusions about the susceptibility of the majority of the American population to implicit prejudice. *See infra* Part III.A.5.

[129] For more on this possibility, see *infra* Part III.B.

be deceptive. There is nothing straightforward about either the theoretical logic or the empirical results.

From one theoretical standpoint, one would expect large correlations between implicit and explicit measures of prejudice. After all, both sets of measures tap into the same target construct of "prejudice." But, from another standpoint, one would expect minimal connections. Recall that the original rationale for the implicit prejudice research program was that there is much to be learned about prejudice by tapping into attitudes that people either cannot or will not consciously endorse when responding to traditional self-report scales. Implicit measures are supposed to be free of social-desirability response biases and to allow us to escape the limits of human introspection, giving us an opportunity to discover attitudes that we hide not just from others but from ourselves. Unfortunately, these opposing theoretical arguments give researchers logical license to adopt a "heads-we-win-tails-we-don't-lose" attitude toward data analysis: they can treat high correlations as evidence for the convergent validity of implicit and explicit measures of prejudice;[130] they can treat low correlations as evidence for the dissociation predicted by dual-process models of cognition—or as just the cumulative result of too much measurement error; they can treat moderate correlations as evidence for all of the above; and they can treat negative correlations as evidence of even more dramatic dissociation between conscious and unconscious cognition, such as repression and reaction formation of the sort posited by now-out-of-fashion psychodynamic theories.

The empirical evidence does little to resolve these uncertainties. In part, this is so because the evidence is so mixed. The median result of studies of the implicit-explicit linkage yield estimates of low positive correlations

---

[130] *See, e.g.*, Cunningham et al., *supra* note 76, at 1342 ("Current research suggests that implicit and explicit attitudes may not be as dissociated as once thought. Whereas initial work showed few correlations between implicit and explicit measures, more recent studies have found correlations between these types of measures." (citations omitted)); Bertram Gawronski, *What Does the Implicit Association Test Measure? A Test of the Convergent and Discriminant Validity of Prejudice-Related IATs*, 49 EXPERIMENTAL PSYCHOL. 171, 178 (2002) ("In the present study, prejudice-related IATs exhibited not only convergent validity to corresponding explicit measures of prejudice endorsement, but also discriminant validity by revealing substantial relations only when the out-group category in the IAT was identical to that in the explicit measure."). Blanton and Jaccard point out a particularly stark example of this theoretical inconsistency. In 2001, Banaji treated low correlations between the IAT and explicit attitude measures as "evidence of validity, not a challenge to it," on grounds that conscious and unconscious attitudes are conceptually distinct, yet just two years later Greenwald, Nosek, and Banaji treated high correlations between the IAT and explicit measures as "evidence for the validity of their new [IAT] scoring algorithm." Hart Blanton & James Jaccard, *Arbitrary Metrics Redux*, 61 AM. PSYCHOL. 62, 66 (2006).

between measures.[131] The preferred interpretation of these results invokes the superiority of implicit measures: such measures are free of the self-deception and self-presentation strategies that contaminate measures of explicit prejudice.[132] An alternative explanation draws on dual-process models of cognition to assign implicit attitudes greater weight in the determination of spontaneous behaviors, such as behavior in reaction-time tests, and to assign explicit attitudes greater weight in the determination of more deliberative behaviors, such as voting choices.[133] These delicate efforts to accommodate

---

[131] *See* Boniecki & Jacks, *supra* note 87, at 11 ("[O]ur research suggests that measures of implicit prejudice and measures of explicit prejudice are generally unrelated."); Brauer et al., *supra* note 53, at 94 (reporting weak to non-existent correlations between various implicit and explicit measures of prejudice); Wilhelm Hofmann et al., *A Meta-Analysis on the Correlation Between the Implicit Association Test and Explicit Self-Report Measures*, 31 PERSONALITY & SOC. PSYCHOL. BULL. 1369, 1379 (2005) (noting that "we found a small but significant positive mean population correlation of .24 between self-reported representations and representations assessed with the IAT" (citation omitted)); Andrew Karpinski & James L. Hilton, *Attitudes and the Implicit Association Test*, 81 J. PERSONALITY & SOC. PSYCHOL. 774, 786 (2001) ("[A]cross all three studies, we consistently failed to find any correlations between the IAT and explicit attitude measures, even when social desirability pressures were minimized."); Brian S. Lowery et al., *Social Influence Effects on Automatic Racial Prejudice*, 81 J. PERSONALITY & SOC. PSYCHOL. 842, 847 (2001) ("Replicating the generally unreliable relation between implicit and explicit attitudes, results reveal that automatic prejudice was uncorrelated with all four measures of explicit attitudes." (citation omitted)); *see also* Fazio & Olson, *supra* note 5, at 303 ("Within the domain of prejudice and stereotypes, the correlations [between explicit and implicit measures] tend to be quite low, although there are occasional reports of significant correlations." (citations omitted)).

[132] *See, e.g.*, Michaël Dambrun & Serge Guimond, *Implicit and Explicit Measures of Prejudice and Stereotyping: Do They Assess the Same Underlying Knowledge Structure?*, 34 EUR. J. SOC. PSYCHOL. 663, 673 (2004) ("[A]n absence of relationship or even a negative relationship between implicit and explicit measures of prejudice is not necessarily incompatible with the conceptualization which argues that implicit and explicit measures are two ways of assessing similar underlying knowledge structure."); *id.* (suggesting that "an 'over-compensation' self-presentational strategy is responsible for the negative relationship between implicit and explicit measures of prejudice"); Hofmann et al., *supra* note 131, at 1369 ("Evidence for the success in assessing meaningful constructs that are difficult to tap with self-reports is implied by the finding that implicit measures often show rather low correlations with explicit measures yet reliably predict behavior." (citations omitted)).

[133] *See, e.g.*, Bertram Gawronski et al., *It's in the Mind of the Beholder: The Impact of Stereotypic Associations on Category-Based and Individuating Impression Formation*, 39 J. EXPERIMENTAL SOC. PSYCHOL. 16, 28 (2002) ("[A]ssociative strength might be a good predictor for spontaneous processes in impression formation such as context effects of category information on the interpretation of ambiguous behavioral cues, whereas explicitly assessed beliefs should be better in predicting processes of deliberate

explicit and implicit prejudice within a single nomological network are upset, however, by the occasional study that finds a substantial correlation between the two types of measures.[134]

The evidence is also inconclusive for a more fundamental reason. The explicit measures of prejudice used in most construct validation studies are themselves politically controversial—and open to alternative explanations. For instance, measures of modern, symbolic, and aversive racism often include items that could equally easily serve as measures of ideological conservatism, traditional values, and the Protestant work ethic.[135] If there are conservatives among implicit prejudice researchers (and there are unlikely to be many in this ideologically-skewed field),[136] they could just as plausibly

---

dispositional inference." (citation omitted)); Andrew Karpinski et al., *Attitude Importance as a Moderator of the Relationship Between Implicit and Explicit Attitude Measures*, 31 PERSONALITY & SOC. PSYCHOL. BULL. 949, 960–61 (2005) ("If one intends to predict a conscious, deliberative behavior and participants are willing and able to report their attitudes on explicit measures, then there is strong evidence that explicit attitudes can predict the behavior and that IAT scores have little predictive value above and beyond explicit attitudes. Conversely, if one intends to predict spontaneous, non-conscious aspects of behavior, then implicit attitudes measures, including the IAT, may be of greater value in predicting the behavior than explicit attitude measures."); Poehlman et al., *supra* note 62, at 27 ("[L]ow correlations between the associations tapped by the IAT and responses on explicit measures were associated with relatively worse predictive validity for both IAT and explicit measures. When implicit-explicit correspondence is low, the associations measured by IAT measures may correlate more weakly with criterion measures because the person is attempting to intentionally override an unwanted automatic response."). *But see* Greenwald & Krieger, *supra* note 76, at 962 (noting that, while the predictive validity of explicit measures is greater than implicit measures for deliberative behaviors, "prediction of behavior by IAT measures was *not* reduced when the examined behavior was more deliberative").

[134] *See* Brauer et al., *supra* note 53, at 84 tbl.3.

[135] *See* Sniderman & Tetlock, *Symbolic Racism*, *supra* note 60, at 130; *see also* Duckitt, *supra* note 33, at 568 ("The concept of new racisms has not been without controversy. Critics have asserted that symbolic racism has been conceptualized and measured inconsistently over time and that the varying themes identified with it have not yet been coherently articulated or adequately measured." (citations omitted)); Brad T. Gomez & J. Matthew Wilson, *Rethinking Symbolic Racism: Evidence of Attribution Bias*, 68 J. POL. 611, 622–23 (2006) ("Our findings . . . suggest that much of the relationship between political sophistication and symbolic racism stems not from racial animus, but from differential patterns of attribution that reach well beyond the domain of race. . . . Since much of the traditional symbolic racism scale is based on respondent attributions of causality, the general tendency of less sophisticated individuals to seize upon localized explanations for sociopolitical events can easily be misconstrued as racial hostility.").

[136] "It is well documented that, like social scientists in general, both academic and practicing psychologists are much more liberal than the general population and most other professionals." Richard E. Redding, *Sociopolitical Diversity in Psychology: The Case for Pluralism*, 56 AM. PSYCHOL. 205, 205 (2001).

insist that the IAT taps into tacit knowledge of depressingly real correlations between race and socioeconomic outcomes in society at large and that the self-report measures of modern forms of racism tap into historically well-grounded recognition of how difficult it will be to solve the societal problems through simple transfer payments. We are not endorsing the conservative counter-interpretation. We are simply asking: How much certainty do we gain from documenting low-positive correlations between measures of dubious validity of one scientifically under-defined construct with measures of dubious validity of another scientifically under-defined construct?

3. *Much Construct Validation Research Employs Weak Criterion Variables Open to Alternative Psychological Explanations and of Questionable Legal Relevance*[137]

A third line of construct validation research explores the power of implicit prejudice measures to predict interpersonal behavior, typically of a subtle nonverbal sort, toward people from varying ethnic-racial groups.[138] As with the prior two types of validation evidence, evidence in this category is not nearly as convincing as advocates of swift application of implicit prejudice research to the law might hope. In fact, one study demonstrated that African-Americans actually preferred to interact with people classified as implicit racists by the IAT.[139] Dramatic disconfirmations of this sort illustrate how easily implicit prejudice researchers can dismiss dissonant evidence.[140] It is not, however, the occasional effect reversal that renders this

---

[137] A criterion variable is an observable behavior or outcome that should correlate reasonably well with a psychological test's scores if the test measures what it purports to measure. For instance, if the LSAT is a good measure of potential for achievement in law school, then it should correlate significantly with law school GPA. *See generally* RALPH L. ROSNOW & ROBERT ROSENTHAL, BEGINNING BEHAVIORAL RESEARCH: A CONCEPTUAL PRIMER 146–47 (3d ed. 1999).

[138] *See, e.g.*, McConnell & Leibold, *supra* note 85, at 438 (noting that subjects who took the IAT and interacted with White and Black experimenters were coded for a variety of indicators of "body language" during the interaction).

[139] *See* J. Nicole Shelton et al., *Ironic Effects of Racial Bias During Interracial Interactions*, 16 PSYCHOL. SCI. 397, 401 (2005) ("Black participants evaluated Whites with higher automatic-bias scores more positively than Whites with lower automatic-bias scores.").

[140] The authors of this study note that Black participants rated high-bias IAT scorers as more engaged. They then argue, with no intended irony, that high-prejudice Whites created better impressions on Blacks because they had to work harder to suppress their prejudice whereas low-prejudice Whites, with less to hide, were more relaxed and came across as disengaged. *See id.* at 401. Of course, one could use the same explanation to account for why, in other studies, allegedly high-prejudice Whites create worse

category of evidence unconvincing for application to the law, for the majority of the evidence here supports the hypothesis that high scorers on implicit measures of prejudice are more likely to exhibit criterion behaviors than low scorers. Rather, most problematic from a legal perspective are the particular behaviors chosen to serve as criterion variables for implicit measures of prejudice.

In most of these "predictive validity" studies, the criterion behaviors involve what Ziegert and Hanges call "micro-level" behaviors, such as nonverbal indicators of interpersonal discomfort and anxiety in the presence of members of the groups toward whom one is allegedly prejudiced or initial impressions about persons of different races.[141] For instance, in one of the most prominent tests of the relationship between scores on the IAT and discriminatory behavior, McConnell and Leibold videotaped subjects who interacted with a White experimenter before taking the IAT and a Black experimenter afterwards and then coded the subjects' behaviors during the interactions for friendliness, comfort level, eye contact, body posture, and other "body language" indicators.[142] They found a significant correlation

---

impressions: efforts to suppress their prejudice make them look nervous. To escape this circularity, it is necessary to entertain possibilities that escaped both the authors' and journal reviewers' attention: (a) all students came from a liberal campus environment and the low scorers may have been so far to the left that Blacks saw them as phony or ingratiating; and (b) high IAT scorers have a more realistic view of the challenges confronting poorer Blacks on largely upper-middle-class campuses and Blacks react favorably to realism. Readers can no doubt generate more possibilities—an exercise in political even-handedness that we also urge for results that cast high IAT scorers in a less flattering light.

[141] *See* Ziegert & Hanges, *supra* note 95, at 553 ("Although research has documented that these implicit measures correlate with other attitudes and predict microlevel behavior, there is currently little evidence indicating that such implicit attitudes are useful for predicting more macrolevel behavior, such as discriminatory hiring decisions."). For a summary of the behavioral/outcome measures employed as criterion variables in the implicit prejudice research, see Dasgupta, *supra* note 36, at 152–55 tbl.I.

Proponents of the IAT tout a recent meta-analysis as showing that "in the context of social group discrimination, implicit attitudes outperform explicit measures in prediction." Banaji et al., *supra* note 29, at 282. Two important qualifications attach to this claim. First, the acts of "social group discrimination" referred to here primarily involve nonverbal and other micro-level behaviors. *See* Poehlman, *supra* note 62, tbl.1. Second, both implicit and explicit measures of prejudice were significant predictors of acts of discrimination in these predictive validity studies, but implicit measures performed better in this domain (r =.25 versus r =.13). *Id.* at 19. However, using Cohen's effect size standards, the effect size for implicit measures in the discrimination domain falls between small to moderate. *See* Kang, *supra* note 16, at 1593 n.110. Further, across all studies, explicit measures of attitudes predicted behavior better than implicit measures (average r =.35 versus average r =.27). Poehlman et al., *supra* note 62, at 21.

[142] McConnell & Leibold, *supra* note 85, at 437–38.

between subjects' IAT scores and ratings of social interaction bias exhibited by subjects: "as participants' IAT scores reflected relatively more positive attitudes toward Whites than Blacks, social interactions were more positive toward the White experimenter than the Black experimenter as assessed both by trained judges and by the experimenters themselves."[143]

Even if we assume for the moment that these differences in interactions with White and Black experimenters do represent deep-seated, unconscious prejudice (and researchers rarely test alternative explanations such as unfamiliarity, guilt, embarrassment, and social anxiety),[144] the legal implications of this research are not apparent. Unless advocates of the implicit prejudice viewpoint favor extraordinarily intrusive state regulation of interpersonal relations—down to eye blinking and postural orientation—as a policy goal in itself, any legal prescription drawn from this research must assume that these micro-level "discriminatory" behaviors accumulate into more consequential discriminatory decisions and actions.[145] For instance, Professor Kang's argument for FCC regulation of news media to alter depictions of minorities in hopes of lessening implicit prejudice assumes that this implicit prejudice leads to consequential actions for minorities and women, in the form of missed job opportunities, poorer educational outcomes, and, most drastically, being shot by the police.[146] Furthermore, the

---

[143] *Id.* at 439.

[144] For instance, the presence of an African-American experimenter "may simply have served to enhance the salience of extrapersonal assocations" (that is, the African-American experimenter may have primed cultural stereotypes or beliefs that were not personally held by the subject that otherwise would not have come to mind). H. Anna Han et al., *The Influence of Experimentally Created Extrapersonal Associations on the Implicit Association Test*, 42 J. EXPERIMENTAL SOC. PSYCHOL. 259, 270 (2006); *see also infra* Part III.A.4.d. For other potential explanations of the effect found in the McConnell and Leibold study, see Chugh, *supra* note 49, at 211–12.

[145] *See* Christine Jolls, *Antidiscrimination Law's Effects on Implicit Bias*, *in* BEHAVIORAL ANALYSES OF WORKPLACE DISCRIMINATION 16 (Mitu Gulati & Michael Yelnosky eds., forthcoming 2006) (manuscript at 16, on file with authors) *available at* http://www.law.yale.edu/documents/pdf/Jolls_Antidiscrimination_Laws_Effects_on_Implict_Bias.pdf ("[E]vidence linking measures of implicit bias to observed behavior does not establish any connection between such measures and *the types of decisions that antidiscrimination law polices*."). Of course, nonverbal behaviors may be important in work settings, but to date the courts and Congress have shown little interest in enforcing civility codes that might reach the level of subtle, nonverbal behaviors. *See, e.g.*, Oncale v. Sundowner Offshore Servs., Inc. 523 U.S. 75, 80 (1998) (rejecting the view that Title VII creates a "general civility code for the American workplace").

[146] Professor Kang weaves together a set of studies showing the potential behavioral consequences of implicit prejudice, but, as Kang discloses, in many of these experiments no measure was taken of participants' implicit bias or implicit prejudice. *See* Kang, *supra* note 16, at 1514–28. Hence, it is pure conjecture that unconscious prejudice caused any

stranger-stranger interactions employed in most social-psychological experiments on discrimination do not reflect the types of interactions that give rise to most employment discrimination lawsuits.[147] Until the connection between measures of implicit prejudice and discriminatory behaviors of greater consequence is established, the claimed link between implicit prejudice and discriminatory behavior as expansively defined by

---

behaviors in those studies. In the few studies he cites where behavior and scores on tests of implicit prejudice are linked, the behaviors are nonverbal or involve hypothetical consequences. Perhaps the most distressing examples Kang cites involve the "shooter bias" studies. In one of these studies, participants under time pressure who were subliminally-primed by a picture of a Black face were more likely to identify incorrectly an ambiguous object as a gun, compared to participants primed by a picture of a White face. *See id.* at 1525; *see also* B. Keith Payne, *Prejudice and Perception: The Role of Automatic and Controlled Processes in Misperceiving a Weapon*, 81 J. PERSONALITY & SOC. PSYCHOL. 181, 187 (2001). In another of these studies, participants under time pressure who were playing a police simulation video game were more likely to shoot unarmed Black targets and refrain from shooting armed White targets. This latter study, however, did not employ implicit measures of prejudice and found no relationship between explicit measures of prejudice and this anti-Black shooter bias. *See* Kang, *supra* note 16, at 1526; *see also* Joshua Correll et al., *The Police Officer's Dilemma: Using Ethnicity to Disambiguate Potentially Threatening Individuals*, 83 J. PERSONALITY & SOC. PSYCHOL. 1314, 1323 (2002). The tendency of African-Americans to display the same bias further undercuts a sweeping prejudice explanation and favors a specific-adaptation-to-local-environments explanation. *See id.* at 1325 ("Testing both White and African-American participants, we found that the two groups display equivalent levels of bias."); *id.* ("The fact that Shooter Bias in Study 3 was related to perceptions of the cultural stereotype, rather than prejudice or personally endorsed stereotypes, suggests that mere knowledge of the stereotype is enough to induce this bias."); *see also* E. Ashby Plant et al., *Eliminating Automatic Racial Bias: Making Race Non-Diagnostic for Responses to Criminal Suspects*, 41 J. EXPERIMENTAL SOC. PSYCHOL. 141, 151 n.4 (2005) (reporting the same pattern of bias across race of participants). Also in this connection, Ashby Plant and her colleagues have shown that repeated exposure to suspects whose race is unrelated to the presence or absence of a gun significantly decreased this anti-Black shooter bias in a police simulation game. *See id.* at 153 (college student participants); E. Ashby Plant & B. Michelle Peruche, *The Consequences of Race for Police Officers' Responses to Criminal Suspects*, 16 PSYCHOL. SCI. 180, 182 (2005) (police officer participants).

[147] *See* David Copus, *A Lawyer's View: Avoiding Junk Science*, *in* EMPLOYMENT DISCRIMINATION LITIGATION: BEHAVIORAL, QUANTITATIVE, AND LEGAL PERSPECTIVES 450, 453 (Frank J. Landy ed., 2005) ("I estimate that upward of 90 percent or more of all employment discrimination lawsuits involve current or former employees, none of whom are strangers to company decision makers."); *id.* at 458 ("So, as long as stereotype research uses stranger-to-stranger interactions in which subjects are given very limited information under conditions that tightly control extraneous independent variables, that stereotype research will have no ecological validity, even if lifelike simulations are used with non-college student participants."); *see also infra* Part III.D.

social psychologists holds little legal significance.[148]

Accordingly, an important avenue for future predictive validity studies from a legal perspective concerns the relationship between implicit associations and subjective evaluations of disadvantaged groups, which may invoke less deliberation in an unstructured setting,[149] and which may be subject to fewer social desirability constraints because ambiguity about ideal qualifications for a position and about the qualifications of a candidate provide cover for individual variation in evaluations.[150] Replicable demonstrations that implicit associations predict bias in the subjective work evaluations of protected groups would bolster legal concerns about subjective hiring and promotion standards.[151]

---

[148] This is particularly true given ambiguity about what exactly causes these behaviors and the robustness of these behavioral effects to situational changes. We consider whether these behaviors unequivocally indicate prejudice or some other social or psychological process in the section on internal validity. *See infra* Part III.B. We consider the generalizability constraints on the relation between implicit prejudice and interpersonal behavior in the section on external validity. *See infra* Part III.D. Assuming these behaviors are the result of whatever psychological process(es) the IAT measures, however, our concern here is whether these behaviors are sufficiently persuasive evidence of discrimination for purposes of legal regulation.

[149] *See, e.g.*, Allen I. Huffcutt & Philip L. Roth, *Racial Group Differences in Employment Interview Evaluations*, 83 J. APPLIED PSYCHOL. 179, 186 (1998) ("We found that group differences for structured interviews tended to be relatively low overall and lower than those for unstructured interviews. One implication of these findings is that structured interviews do tend to limit the influence of biases and stereotypes."); Joshua M. Sacco et al., *An Investigation of Race and Sex Similarity Effects in Interviews: A Multilevel Approach to Relational Demography*, 88 J. APPLIED PSYCHOL. 852, 860 (2003) ("Our results suggest that organizations do not have to be concerned about matching interviewer and interviewee race and sex to avoid potentially biased ratings, at least when using carefully developed job-analysis-based highly structured interviews like the ones used here."); Laura Gollub Williamson et al., *Employment Interview on Trial: Linking Interview Structure with Litigation Outcomes*, 82 J. APPLIED PSYCHOL. 900, 908 (1997) ("Past research has shown that structuring the interview can be a key to enhancing its reliability and validity. The purpose of this study was to test the thesis that structuring can also be a key to enhancing its legal defensibility. Two approaches were taken to test this thesis, one conceptual and one empirical. Both were strongly supported.").

[150] *See, e.g.*, Lu-in Wang, *Race as Proxy: Situational Racism and Self-Fulfilling Stereotypes*, 53 DEPAUL L. REV. 1013, 1018 (2004) ("[I]ndividuals are most likely to discriminate in situations in which their behavior is least likely to be viewed as discriminatory—thereby providing 'cover' for their racially biased conduct." (emphasis omitted)).

[151] *See* Melissa Hart, *Subjective Decisionmaking and Unconscious Discrimination*, 56 ALA. L. REV. 741, 748 (2005). Existing evidence, however, casts considerable doubt on the assumption that subjective measures lead to greater bias in employment actions than objective measures. *See* H.W. Hennessey, Jr. & H. John Bernardin, *The Relationship*

To date, only studies by Ziegert and Hanges and by Rudman and Glick directly address this question, and these studies provide far-from-decisive evidence that implicit associations are pervasive and potent causes of discrimination in a subjective candidate or employee evaluations. In a workplace simulation, Ziegert and Hanges found a relation between implicit racial associations and discriminatory ratings of Black job candidates, but only when the "boss" explicitly told subjects to engage in racial discrimination[152] (and, even then, only when the data analysis included a handful of anti-Black outliers on the dependent variable).[153] Rudman and Glick found a relationship between subjective judgments of the social skills of an "agentic" (as opposed to a "communal") female applicant for a job requiring social sensitivity and subjects' scores on a version of the IAT designed to measure gender stereotypes, but not between subjects' subjective judgments of the "hireability" of this applicant and IAT scores (i.e., there seemed to be a stronger "backlash" against agentic, counter-stereotype

---

Between Performance Appraisal Criterion Specificity and Statistical Evidence of Discrimination, 42 HUM. RES. MGMT. 143, 156 (2003) ("We must conclude that the burden is certainly on those experts who maintain that there is some causal connection between a particular deleterious outcome for protected class member(s) and a particular type of performance appraisal format or system. Our data do not support the argument."); Philip L. Roth et al., *Ethnic Group Differences in Measures of Job Performance: A New Meta-Analysis*, 88 J. APPLIED PSYCHOL. 694, 702 (2003) ("Our results do not support the position that subjective measures have more potential for bias than objective measures. Instead, we found the opposite.").

[152] Ziegert and Hanges demonstrated a relation between IAT scores and discriminatory hiring recommendations, but only in a "climate for racial bias" in which the president of a hypothetical company had written a memo expressly directing managers (played by the subjects) to favor White candidates over Black candidates. *See* Ziegert & Hanges, *supra* note 95, at 558 (The memo from the president to the manager began with this sentence: "Given that the vast majority of our workforce is White, it is essential we put a White person in the VP position."). This memo would almost certainly constitute direct evidence of discriminatory intent under current employment discrimination law. *See, e.g.*, Rudin v. Lincoln Land Cmty. Coll., 420 F.3d 712, 720 (7th Cir. 2005) ("Direct evidence 'can be interpreted as an acknowledgment of discriminatory intent by the defendant or its agents.'"); KEVIN F. O'MALLEY ET AL., FEDERAL JURY PRACTICE AND INSTRUCTIONS CIVIL § 170.20 (2005) ("Direct evidence would include statements showing a discriminatory motivation for the defendant's treatment of the plaintiff."). Indeed, in light of this memo directing the manager to discriminate—a tremendous demand effect—it would have been surprising had Ziegert and Hanges not found greater evidence of discrimination in this condition. The study also asked subjects playing the role of managers to evaluate applicants on the basis of paper dossiers alone, Ziegert & Hanges, *supra* note 95, at 556, a practice which may occur with some regularity but which rarely gives rise to discrimination lawsuits. *See* Copus, *supra* note 147, at 452 ("[E]mployment discrimination lawsuits challenging alleged discrimination in the screening of resumes are exceptionally rare.").

[153] *See infra* Part III.C.2.

women in social skills ratings by subjects who scored higher on the gender stereotype IAT, but this relationship between subjective judgments and IAT scores was not found with respect to hireability ratings).[154]

A few other studies examine the relation between implicit prejudice and subjective evaluations of social group members, but not in employment settings, which present a host of special external validity concerns.[155] Ashburn-Nardo and colleagues examined the relation between African-Americans' scores on the race IAT and their judgments about a Black or White person as a partner on an upcoming anagram-solving task.[156] They found no relation between subjects' actual choices of partners and IAT scores,[157] but they did find a significant correlation between IAT scores and subjects' expectations and attitudes about partnering with a White or Black person.[158] Rudman and Lee found a correlation between subjects' scores on a race stereotype IAT and subjects' ratings of the hostility and sexism evidenced by a Black target's ambiguous behavior, but this correlation did not hold for subjects' ratings of the Black target's intelligence or for any ratings of a White target.[159]

Given the mixed evidence on implicit measures of prejudice as predictors of subjective judgments, a priority for would-be exporters of IAT findings to the law should be to examine the relationship between IAT scores and subjective evaluations of men, women, and minorities for suitability in different jobs and for performance in different jobs, and the moderators of

---

[154] *See* Laurie A. Rudman & Peter Glick, *Prescriptive Gender Stereotypes and Backlash Toward Agentic Women*, 57 J. SOC. ISSUES 743, 756–57 (2001).

[155] For a discussion of external validity considerations when extending basic research to employment settings, see *infra* Part III.D.

[156] Leslie Ashburn-Nardo et al., *Black Americans' Implicit Racial Associations and Their Implications for Intergroup Judgment*, 21 SOC. COGNITION 61, 66 (2003).

[157] *Id.* at 70 n.2.

[158] *Id.* at 75 ("The more blacks implicitly preferred their own race relative to whites, the greater their preference for a black relative to a white work partner."). Note that the researchers used the phrase "partner preference" to refer to judgments about the White and Black potential partners, rather than to refer to the actual choice of partners. Note also that this correlation was only marginally significant after controlling for explicit measures of racism. *Id.* at 76.

[159] Laurie A. Rudman & Matthew R. Lee, *Implicit and Explicit Consequences of Exposure to Violent and Misogynous Rap Music*, 5 GROUP PROCESSES & INTERGROUP RELATIONS 133, 143 (2002); *see also* Gawronski et al., *supra* note 133, at 26 (using a median split of scores on a gender stereotype IAT to group people as possessing strong or weak gender stereotype associations and finding that strong stereotype associations were correlated with a gender bias in target attributions when guessing whether a man or woman made a statement).

this relationship.[160] However, these studies should—at a minimum—control for objective differences in job performance across groups, which may reflect true differences in performance or may reveal that objective measures have even greater adverse impact than subjective evaluations.[161]

4. *Alternative Psychological Explanations of the IAT Effect*

Some legal scholars carefully avoid labeling high scores on the IAT evidence of unconscious prejudice.[162] Recent research underscores the wisdom of such circumspection. There is strong evidence that psychological processes aside from out-group hostility can artificially inflate and otherwise distort scores on implicit measures such as the IAT.[163] These alternative processes include the power of stimulus familiarity/unfamiliarity to facilitate/impede reaction time, the power of compassion to increase the accessibility of negatively-charged cognitions, the power of widely known but not personally accepted cultural stereotypes to influence the accessibility of associations, the power of objective correlations between group membership and socio-economic outcomes to influence the accessibility of associations, and the power of individual differences in cognitive flexibility to influence reaction time in response to shifting instructions such as those employed with the IAT.

a. *Figure-Ground Asymmetry, Not Antipathy*

One challenger to the implicit-prejudice explanation of the IAT effect is

---

[160] *See* Dasgupta, *supra* note 36, at 157–58 ("Just as implicit attitudes have, in recent years, been shown to be remarkably malleable, so too behaviors are also quite malleable depending on the extent to which awareness, control, and motivation are at play." (citation omitted)).

[161] *See generally* Roth et al., *supra* note 151, at 702; Frank L. Schmidt & John E. Hunter, *The Validity and Utility of Selection Methods in Personnel Psychology: Practical and Theoretical Implications of 85 Years of Research Findings*, 124 PSYCHOL. BULL. 262 (1998).

[162] *See, e.g.*, Eisenberg & Johnson, *supra* note 108, at 1551 (indicating that "when we say that subjects have an 'automatic preference for white,' we mean nothing more than that they automatically associate white with good and black with bad").

[163] *See, e.g.*, Frederica R. Conrey et al., *Separating Multiple Processes in Implicit Social Cognition: The Quad Model of Implicit Task Performance*, 89 J. PERSONALITY & SOC. PSYCHOL. 469, 483 (2005) ("[O]ur findings suggest that researchers should exercise caution in assuming that the implicit prejudice scores they calculate with priming measures or with the IAT reflect exclusively the strength of automatic associations.").

the "figure-ground model."[164] In this model, IAT effects do not carry the morally loaded meaning assigned them by IAT proponents (unconscious negative evaluations of minorities) because the effects can be better explained by confounding variation in the salience of the specific stimuli used in the test.

The figure-ground model starts from the widely accepted assumption among cognitive psychologists that the human brain organizes visual experience into figure-ground contrasts: some things occupy the foreground in consciousness (the "figure") and other things recede into the background (the "ground"). People partition their experience so that the novel, strange, and unexpected stand out against the routine, normal, and expected.[165] Accordingly, we should generally expect members of out-groups to stand out against the more routine backdrop of one's in-group:

> [D]ifferences in salience reflect an even more fundamental distinction than differences in evaluation. Figure-ground asymmetries are closely related to the regulation of behavior. Confronting a stimulus of a figure category causes an allocation of attention, controlled processing, an interruption of ongoing processing and behavioral routines, and a reorienting of cognition and action. The capacity of salient information to capture processing and

---

[164] Rothermund and Wentura call this explanation the "[f]igure-[g]round [m]odel of IAT [e]ffects," Klaus Rothermund & Dirk Wentura, *Underlying Processes in the Implicit Association Test: Dissociating Salience From Associations*, 133 J. EXPERIMENTAL PSYCHOL.: GENERAL 139, 140 (2004), while Kinoshita and Peek-O'Leary refer to it as the "salience asymmetry account," Kinoshita & Peek-O'Leary, *supra* note 26, at 444. "The main tenet of this account is that the IAT effect is driven by asymmetries in salience between the contrasting categories." *Id.* Proctor and Cho, in turn, refer to the IAT effect as an example of the "salient features coding principle" at work: "The stimulus and response sets are coded with respect to their salient features, and translation of a stimulus into a response is fastest for mappings in which the salient features of the sets correspond." *See* Robert W. Proctor & Yank Seok Cho, *Polarity Correspondence: A General Principle for Performance of Speeded Binary Classification Tasks*, 132 PSYCHOL. BULL. 416, 418 (2006). Proctor and Cho note that this account has the advantage of parsimony over the evaluation account because: (1) it fits the data from a wide range of speed binary classification tasks, of which the IAT is a special case, *see id.* at 418; and (2) it is possible to obtain IAT effects even when the attribute categories do not differ in valence. *See id.* at 435.

[165] Recently Nisbett and his colleagues have shown that East Asians tend to attend to the environment as a whole while Westerns tend to focus on particular objects within the environment. Li-Jun Ji et al., *Culture, Control, and Perception of Relationships in the Environment*, 78 J. PERSONALITY SOC. & PSYCHOL. 943, 952 (2000); Takahiko Masuda & Richard E. Nisbett, *Attending Holistically Versus Analytically: Comparing the Context Sensitivity of Japanese and Americans*, 81 J. PERSONALITY SOC. & PSYCHOL. 922, 933 (2001). Therefore, disparity in attention to the figure relative to the ground may be culturally determined rather than a universal feature of cognition.

> action resources, however, is not tied to a specific valence. For example, both positive and negative other-relevant traits (e.g., generous, aggressive) automatically attract attention and interrupt current behavioral routines irrespective of their valence.[166]

The figure-ground model attributes the IAT effect to a confounding variable: the failure to control for asymmetries in salience between the contrasting categories. When subjects confront the taxing binary-classification task used in the IAT (whether to press the left or right key when shifting combinations of pleasant-versus-unpleasant and young-versus-old stimuli flash up on the screen), they try to simplify that task by focusing on only one of the contrasting categories.[167] People reflexively place the more unusual category in the perceptual foreground—and relegate the other category to the background. This leads to the prediction that performance should be facilitated if people are asked to press the same key (say, left) to both of the stimuli in the perceptual foreground and the same key to both of the stimuli in the perceptual background. Researchers have found exactly such facilitation when they manipulated the focal category by instructing subjects to respond to only one category in the individual target and attribute classification blocks (e.g., "respond if the item is a young [old] person's name and make no response otherwise").[168]

Brendl and colleagues also demonstrate the power of figure-ground asymmetries by tweaking the original demonstration of the IAT as a measure of implicit attitudes. In the first published report on the IAT, Greenwald and colleagues constructed a flower-insect IAT to show that people more strongly associate flowers with positive words and insects with negative words—and the data seemed to support that contention.[169] Brendl and colleagues, however, substituted nonsense syllables for flowers as one of the target categories and showed that people responded faster when *insects and pleasant words* were assigned one response (say, the left key) and affectively-neutral nonsense syllables (e.g., WAB, ZIL) and unpleasant words were assigned the other response (say, the right key).[170] If we accept the logic underlying Greenwald et al.'s interpretation of the IAT effect, we are now obliged to conclude from the Brendl group's results that people have an implicit *positive* attitude toward insects.

But this leads to a contradiction for the implicit-attitude explanation. If

---

[166] Rothermund & Wentura, *supra* note 164, at 159 (citation omitted).

[167] For a full description of procedure involved in the IAT, see *supra* Part II.B.

[168] Kinoshita & Peek-O'Leary, *supra* note 26, at 444.

[169] *See* Greenwald et al., *supra* note 5, at 1466–70 (Experiment 1).

[170] Brendl et al., *supra* note 71, at 767–68.

the IAT exclusively measures evaluative associations and if the original IAT result reflects an insect-negativity association, then this same association should have prevailed in the nonsense-syllable IAT. By contrast, the figure-ground model readily explains these findings. Nonsense syllables carry, by definition, minimal good-bad meaning but these weird combinations of letters are unfamiliar and thus more distinctive than words that denote everyday insects. The figure-ground asymmetry in the target dimension was thus reversed, and nonsense syllables and negative words formed the figures against a background of positive words and insects. The figure-ground model thus predicts faster responses when insects and positive words are assigned the same response.[171]

Such experiments have led some scholars to conclude that greater familiarity with one ethnic-racial group (e.g., Whites over Blacks) drives at least part of the race IAT effect.[172] Here it is worth emphasizing that this

---

[171] Rothermund and Wentura explain the result thus:

> Assume ... that the two categories of both the target and the attribute dimension indeed differ in salience (e.g., assume that insects are more salient than flowers and that unpleasant words are more salient than pleasant words). In this case, participants would find it easier to respond if the salient categories of both dimensions (the figures) were mapped onto one response and the non-salient categories (the background) were mapped onto the other response. With such a consistent assignment of salient and non-salient categories to responses, category salience helps to discriminate between responses. The categorization task now resembles a simple visual search task with a display set size of one stimulus .... If a stimulus belongs to a salient category, a "*yes*" response is executed .... If the stimulus does not belong to a salient category, a "*no*" response is executed .... In incompatible blocks, by contrast, the salience and response dimensions are ... not mapped consistently onto responses .... [And] there is [therefore] no facilitative influence of salience on responding. The figure-ground model thus assumes that IAT effects reflect independent salience asymmetries within the target and attribute dimensions.

Rothermund & Wentura, *supra* note 164, at 140 (citation omitted).

[172] For instance, Brendl and colleagues caution against a "prejudice only" interpretation of the IAT effect:

> The present studies show that each one of the following people would be diagnosed as negatively prejudiced against African Americans in the Black-name–White-name IAT: A person with (a) prestored negative evaluations of Black names, (b) prestored positive evaluations of White names without evaluative associations to Black names, (c) stronger prestored positive evaluations of White than Black names without negative evaluations of either, and (d) low familiarity of Black names in the absence of any prestored evaluation of Black names. In our view there is little doubt that according to commonly agreed on definitions of prejudice only the first person would be called prejudiced. Whereas prejudice leads to an IAT effect, an IAT effect does not unambiguously indicate prejudice, because it can have multiple causes.

argument, like later arguments in this subsection, need not be "either/or." It is damaging to the "prejudice" interpretation of the IAT effect even if only modest percentages of the variance in IAT scores (say, 10% or 15%) are attributable to salience asymmetries. Such damage can be cumulative in an especially theory-threatening way in light of two additional facts: (a) psychometric estimates of error variance (unreliability) in implicit measures often exceed 50% of the total variance,[173] so conceding any portion of the remaining systematic variance to other theoretical perspectives is troubling to the implicit prejudice account; (b) salience asymmetry is but one of many counter-interpretations of what the IAT and related instruments tap, which we address shortly. Insofar as these additional counter-interpretations of the race IAT effect also explain significant fractions of the variance, it becomes increasingly difficult for defenders of the race IAT not only to portray their instrument as a pure measure of implicit racial attitudes, but even to portray their instrument as primarily a measure of implicit racial attitudes.

One worrisome sign of politicized research programs—in which hypothesis advocacy has supplanted hypothesis testing—is the haste with which counter-interpretations are dismissed. Implicit prejudice advocates confidently assert that they have ruled out familiarity as an explanation of IAT results.[174] If, by this claim, they mean "for all IAT results," the claim is

---

Brendl et al., *supra* note 71, at 768–69; *see also* Kinoshita & Peek-O'Leary, *supra* note 26, at 446 (reviewing evidence "suggestive of the possibility that at least some part of the prowhite IAT effect reflects salience asymmetry, rather than more positive evaluation of whites than blacks").

[173] *See* Cunningham et al., *supra* note 126, at 169 ("Our analyses of implicit attitude measures suggest that the degree of measurement error in response-latency measures can be substantial—estimates of Cronbach's alpha indicated that, on average, more than 30% of the variance associated with the measurements was random error."). Cunningham et al. report alpha coefficients of .78 and .63 for the IATs they studied. *Id.* at 166. If the mean alpha for the IAT is taken to be .70, this yields a rough estimate of error variance of over 50% (using the simple formula of error variance = 1 – alpha coefficient squared). Furthermore, data gathered by Blanton and his colleagues indicates that "the largest source of variation in the component parts of these measures probably is systematic error and that variation due to processing speed sometimes may reflect meaningful variation rather than simple method variance." Hart Blanton et al., *Decoding the Implicit Association Test: Implications for Criterion Prediction*, J. EXPERIMENTAL SOC. PSYCHOL. (forthcoming 2006) (manuscript at 27, on file with authors).

[174] Nilanjana Dasgupta et al., *Automatic Preference for White Americans: Eliminating the Familiarity Explanation*, 36 J. EXPERIMENTAL SOC. PSYCHOL. 316, 325 (2000) ("The present findings refute the familiarity explanation by demonstrating automatic White preference even when subjective familiarity with Black and White exemplars is equated."); *see also* Banaji et al., *supra* note 29, at 282 (relying on the Dasgupta et al., *supra*, study for the proposition that "mere familiarity' cannot account for implicit attitudes measured by the IAT"); Kang & Banaji, *supra* note 3, at 1072 n.46 (claiming that the IAT measures are not confounded with familiarity).

demonstrably false. If, by this claim, they mean that there are at least some IAT results that are difficult to explain in terms of figure-ground asymmetries, they may be right, but the evidence they invoke even for this limited claim is far from decisive.[175] For instance, IAT advocates regularly

---

[175] Rothermund and Wentura recently re-affirmed their position that the figure-ground explanation remains strong despite the efforts of Greenwald and others to undercut it. *See* Anthony G. Greenwald et al., *Validity of the Salience Asymmetry Interpretation of the Implicit Association Test: Comment on Rothermund and Wentura (2004)*, 134 J. EXPERIMENTAL PSYCHOL.: GENERAL 420 (2005); Klaus Rothermund et al., *Validity of the Salience Asymmetry Account of the Implicit Association Test: Reply to Greenwald, Nosek, Banaji, and Klauer (2005)*, 134 J. EXPERIMENTAL PSYCHOL.: GENERAL 426 (2005).

Indeed, in this exchange, Greenwald and colleagues make critical concessions. They acknowledge that: (a) they have worked with a very loose definition of mental "association" that could reflect the operation of a wide range of psychological processes, including frequency, contiguity (in time or place), similarity, contrast, or causation, *see* Greenwald et al., *supra*, at 421; and (b) "salience asymmetries have the potential to contribute to IAT effects," *id.* at 420. Greenwald and colleagues do not, however, surrender. They counter-attack. On the theoretical front, they try to neutralize the salience counter-interpretation by attributing the dispute simply to rival definitions and stipulating that their definition of "association" is broad enough to cover many studies inspired by the rival figure-ground account. Rothermund, Wentura, and De Houwer reply pointedly that the broad conception of association preferred by Greenwald et al. renders the concept of association "meaningless and inconsequential." Rothermund et al., *supra*, at 427. They continue:

> When used in a broad sense, [association] can refer to any process or relation. Also, if an IAT effect is found, one can always find some feature that the concepts of the IAT have in common. Hence, we agree with the view of Kinoshita and O'Leary: "In the absence of a commitment to a theory of association, association can be defined only post hoc, by the data. Thus, the target and attribute dimensions are said to be associated when there is an IAT effect, and unassociated when no IAT effect is observed . . . . This is clearly circular, and the claim that the IAT effect taps associations between the target and attribute dimensions can never be falsified."

Id. at 427. On the methodological front, Greenwald et al. try to neutralize data that support the figure-ground model by arguing that critics have failed to follow standard IAT protocols. Displaying uncharacteristic sensitivity to the problems that arise from using artificial laboratory stimuli, they declare that "the IAT does not function properly" when researchers use words versus non-words as task categories or when researchers fail to use their preferred scoring rules. Greenwald et al., *supra*, at 423. But these objections have no grounding in psychological theory (IAT advocates had already conceded they did not have a theory of response mechanisms) and, if heeded, would prevent researchers from performing the IAT experiments necessary to pinpoint underlying processes (the only way to determine what should constitute the "standard protocol" is by testing non-standard variations). In short, the IAT research program, at the tender age of eight years, is displaying signs of "Lakatosian" degeneration into tautology. *See* Imre Lakatos, *Falsification and the Methodology of Scientific Research Programmes*, *in* CRITICISM AND

cite a study by Dasgupta and her colleagues as definitive evidence against the familiarity explanation.[176] That study used photographs of non-famous Black and White people as stimuli in the IAT and rested on the precarious assumption that these photographs would be equally unfamiliar to research participants.[177] When the IAT revealed faster response times to the pairing of White faces with pleasant words, the authors argued that differential familiarity could not account for this result because pictures of non-famous White and Black persons must be equally unfamiliar.[178] As Kinoshita and Peek-O'Leary point out, this argument ignores the well-replicated "other race effect": people more readily recognize the faces of strangers from their own race than from others.[179] Substantial evidence suggests that this effect is due to familiarity, including work showing that faces from familiar racial groups are recognized more holistically and evidence that the other-race effect can be reduced by frequent cross-racial contact. Similar indeterminacy problems arise with respect to the other lines of evidence that IAT defenders invoke against the familiarity hypothesis, including both brain-imaging and behavioral evidence.[180]

It is not unusual for competing theoretical accounts to coexist for findings from the same experimental task in cognitive psychology. But it is unusual for the societal stakes to be so large and so contingent on which account is correct. From the available evidence, it is reasonable to conclude that: (a) the familiarity account does explain significant fractions of the

---

THE GROWTH OF KNOWLEDGE 91, 118 (Imre Lakatos & Alan Musgrave eds., 1970) (describing "problemshifts" that do not involve novel factual predictions or empirical corroboration of such predictions as signs of degeneration within a research program); *see also id.* at 175 n.2 (describing theoretical moves that predict no novel facts as "*ad hoc* theories"). The growing recognition in the scientific community of the failure of IAT advocates to advance a falsifiable process model should serve as an instructive warning to the legal community.

[176] *See* Dasgupta et al., *supra* note 85, at 240; *see also* Kinoshita & Peek-O'Leary, *supra* note 26, at 446 (noting that this Dasgupta et al. study is "often cited as ruling out familiarity as the basis of the race IAT effect").

[177] The report of this study does not indicate that subjective familiarity with the pictures was pre- or post-tested. *See* Dasgupta et al., *supra* note 174.

[178] *See id.* at 325 ("The present findings refute the familiarity explanation by demonstrating automatic White preference even when subjective familiarity with Black and White exemplars is equated."). The authors of the study also base this conclusion on a race IAT involving "White names" and "Black names" in which they statistically controlled for stimulus familiarity, *see id.* at 322–24, but this study is also problematic for reasons Kinoshita and Peek-O'Leary point out. *See* Kinoshita & Peek-O'Leary, *supra* note 26, at 446–47.

[179] *See id.* at 446.

[180] *See id.* at 446–47 (discussing problems with studies claiming to have ruled out familiarity as an explanation of the IAT effect).

variance in those studies in which it has been directly pitted against the evaluation account; (b) the vast majority of IAT research has not been properly designed if the goal is to determine the precise processes that underlie IAT effects;[181] and (c) given that salience and valence are often naturally intertwined, it will be difficult, perhaps impossible, to disentangle these explanations using the naturalistic stimulus materials preferred by IAT researchers who seek to engage the broader societal debates about racism, sexism, heterosexism, and ageism.

b. *Fear of Being Labeled a Bigot, Not Bigotry*

Implicit measures of prejudice are generally more subtle than the explicit measures of prejudice they have been advanced to complement.[182] But, as anyone who has ever taken the race-based IAT appreciates, it soon becomes obvious that the researchers are curious about how one thinks about race, and this realization can make subjects apprehensive.[183] Consistent with this concern, a recent study found that the more threatened research participants felt by the IAT—the threat of being stereotyped as White racists—the "worse" their scores were on the IAT.[184] "If individuals are highly motivated to control prejudiced responses, then leading them—or allowing them—to

---

[181] *See* Rothermund & Wentura, *supra* note 164, at 159 ("Standard IATs should be accompanied by a corresponding word-nonword version of the task (or by any other technical version of the task that makes use of an asymmetrical attribute dichotomy that is clearly not associated with the target categories).").

[182] For a discussion of methods to use to ameliorate social desirability bias in responses to explicit measures of prejudice, see David O. Sears, *A Perspective on Implicit Prejudice from Survey Research*, 15 PSYCHOL. INQUIRY 293, 295–96 (2004).

[183] Indeed, the web version of the IAT, through which a massive amount of IAT data has been collected, *see, e.g.*, Nosek et al., *supra* note 6, at 102, expressly advises potential participants that various versions of the IAT are meant to assess attitudes toward different groups and indicates that the IAT often reveals preferences for majority groups. For instance, the page on which participants choose a particular IAT test informs the participant that the gender-science IAT "often reveals a relative link between liberal arts and females and between science and males," that the race IAT "indicates that most Americans have an automatic preference for white over black," that the sexuality IAT "often reveals an automatic preference for straight relative to gay people," that the weight IAT "often reveals an automatic preference for thin people relative to fat people," and that the Arab-Muslim IAT "frequently reveals an automatic preference for other people compared to Arab-Muslims." Implicit Association Test, Select a Test, https://implicit.harvard.edu/implicit/demo/selectatest.jsp (last visited Aug. 30, 2005).

[184] Frantz et al., *supra* note 28, at 1621 ("[T]hreatened participants (i.e., those who were told or who guessed what the IAT assessed) had higher IAT effects than nonthreatened participants (i.e., those who did not know the nature of the measure)").

believe that the IAT measures racism will result in elevated IAT effects."[185] These researchers highlight a measurement dilemma: "the race IAT's purpose is extremely difficult to mask and that failing to mask its purpose results in elevated scores."[186]

This feature of the IAT has the potential to create a self-fulfilling prophecy. Implicit prejudice researchers attribute stubbornly persistent inequalities in the U.S. to stubbornly persistent prejudice among White Americans—and develop a test that, while less obvious than most explicit measures of prejudice, is still a fairly transparent gauge of racial sentiment. Many Whites—including the most politically correct among them—then react to the identity threat posed by the IAT by choking under stress—and performing even worse on the IAT, thus confirming the researchers' original stereotype of them.[187]

c. *Sympathy, Not Antipathy*

The alternative explanations considered so far—familiarity confounds and evaluation-apprehension confounds—raise the possibility that IAT effects are *not* driven by negatively-charged mental associations. The next three alternative explanations—sympathy versus antipathy, awareness versus endorsement of cultural stereotypes, and awareness versus justification of depressing realities—concede that some form of associationist account is needed to explain at least some fraction of the systematic variance in IAT scores. But these explanations challenge the characterization of associations as tapping into racial attitudes—at least attitudes in the commonsense view that the attitudes imply an evaluative preference that, when brought to people's attention, they endorse and are even prepared to justify under appropriate conditions.[188]

---

[185] *Id.* at 1622.

[186] *Id.*

[187] *See id.* ("Ironically, the IAT appears to be the most threatening to people who most want to appear nonracist . . . . [T]he present results indicate that under situations of threat, the desire to behave in an egalitarian manner leads to larger, not smaller, IAT effects.")

[188] Implicit prejudice researchers use "attitude" and "association" interchangeably. But Fiedler and colleagues note that equating attitude and association "is problematic because attitudes are one-dimensional" approach-avoidance constructs whereas associations are "multidimensional." Klaus Fiedler et al., *Unresolved Problems with the "I", the "A" and the "T": Logical and Psychometric Critique of the Implicit Association Test (IAT)*, EUR. REV. SOC. PSYCHOL. (forthcoming 2006) (manuscript at 12). Consider four examples: (1) associations can bind synonyms (friend-ally) or antonyms (friend-foe); (2) associations can cause assimilation (negative stereotype of group x can lead to

The "sympathy rather than antipathy" explanation poses the possibility that the allegedly negative associations revealed by the IAT are rooted more in compassion or guilt about the predicament of African-Americans than in hostility or contempt. Consider this thought experiment advanced by Arkes and Tetlock.[189] Imagine two citizens who have very different views on the obstacles to racial equality. One citizen is sympathetic to a left-liberal political agenda, believes that racial discrimination is an ongoing problem, supports affirmative action, and believes that African-Americans are disadvantaged in America today due to the historical legacy of slavery, continuing exploitation, and de facto segregation. This citizen believes that progress during the past half century has been too little too late, and he is convinced that many Whites unfortunately continue to harbor resentment of African-Americans. The other citizen is sympathetic to the right-conservative political agenda and believes in the power of individual responsibility and competition to propel African-Americans up the ladder of success. This citizen disapproves of affirmative action and believes that the primary cause of African-American economic and educational inequality in America today is internal to the African-American community: the widespread abdication of personal responsibility within inner city communities and the surge in the late twentieth century of out-of-wedlock births.

Although the two individuals disagree profoundly on key political issues, they agree on basic facts. They agree that the African-American family is in trouble, that African-American crime rates are far too high, and that African-American educational test scores are too low. They react to these facts with distinct mixtures of sorrow, frustration, and anger directed at distinct political targets.

Is there any compelling theoretical reason to expect these two individuals to exhibit different reaction times on implicit measures or any compelling methodological reason to expect implicit measures, such as the IAT, to differentiate people who share a large knowledge base but differ in their

---

negative evaluations despite surprisingly good performance) or contrast effects (negative stereotype can lead to even more positive evaluations in the wake of surprisingly good performance); (3) associations can be asymmetric (the expected probability of cheddar given cheese is greater than that of cheese given cheddar); and (4) associations can be influenced by superficial factors such as word frequency or rhyming that are independent of semantic overlap. If the implied equivalence of attitudes and associations collapses, so too does the argument for treating implicit associative measures as implicit measures of attitudes and hence prejudice.

[189] *See* Arkes & Tetlock, *supra* note 29, at 257. This article is subtitled, "Would Jesse Jackson 'Fail' the [IAT]?," because Jesse Jackson is at real theoretical risk of failing the IAT given his admission that he associates Blacks more strongly with violent crime than he does Whites. *See id.* at 257, 264–65.

causal attributions for inequality?[190] Promoters of the IAT have presented no evidence that current reaction time measures can reliably distinguish qualitatively different cognitive-emotional states, such as frustration, sorrow and anger rooted in competing cognitive appraisals of the political scene. Indeed, there is suggestive evidence to the contrary. Uhlmann and his colleagues experimentally created fictitious groups, some of which were portrayed as victims of oppression not dissimilar to that experienced by many African-Americans, and found that subjects had more "negative" IAT scores toward the groups who were treated manifestly unfairly and who, most importantly, elicited sympathy, not contempt, from observers.[191]

This alternative sympathy-based account of the IAT effect is every bit as well-positioned as racial-animus accounts to explain the ostensibly "hardest-science" evidence for the construct validity of the IAT, namely, work on neuropsychological correlates.[192] Phelps and colleagues reported that

---

[190] Defenders of the IAT can concede that they have no theoretical basis for distinguishing the two individuals in the thought experiment but still insist they have some empirical basis. The low positive correlations between the IAT and so-called modern racism scales suggest at least limited power to make such distinctions. Putting to the side whether people should be labeled racists for believing that many or most obstacles to racial equality are now internal to the Black community, the unstable and often weak correlational links here mean that substantial percentages of White Americans are failing the IAT even though their belief system is much closer to the first than to the second individual. Full disclosure on the part of IAT advocates should require that they estimate these percentages when they make claims about the pervasiveness of implicit prejudice.

[191] *See* Uhlmann et al., *supra* note 27, at 494; *see also* Nilanjana Dasgupta & Luis M. Rivera, *From Automatic Antigay Prejudice to Behavior: The Moderating Role of Conscious Beliefs About Gender and Behavioral Control*, 91 J. PERSONALITY & SOCIAL PSYCHOL. 268, 277 (2006) ("[T]he present studies are the first to show that, although spontaneous behavior toward stigmatized others may be driven by automatically activated prejudice under some conditions, conscious processes such as the motivation to be egalitarian and behavioral control can circumvent the effect of automatic prejudice on outward behavior."); Gordon B. Moskowitz et al., *Preconscious Control of Stereotype Activation Through Chronic Egalitarian Goals*, 77 J. PERSONALITY & SOC. PSYCHOL. 167, 181–82 (1999) ("Activation of stereotypes can be controlled by more dominant responses, in this case, egalitarian goals. However, this requires commitment to such a goal, and without commitment, stereotype activation will be likely to occur.... Commitment to egalitarian goals can prevent stereotype activation when making inferences from social information.").

[192] Kinoshita and Peek-O'Leary also show how the neuroscience evidence can be explained in their salience-asymmetry framework. *See* Kinoshita & Peek-O'Leary, *supra* note 26, at 447–48. Contrary to claims that we can now read minds for evidence of racism in the brain, there is growing concern among neuroscientists that, although careful research can identify brain sites responsible for key processes such as visual pattern recognition, spatial-mathematical reasoning and language comprehension, we make an

differences in strength of amygdala activation to African-American versus White faces was correlated with bias detected on the IAT.[193] The authors announced that "we have for the first time related indirect behavioral measures of social evaluation to neuronal activity."[194] The authors noted that, while the amygdala is involved in signaling the presence of stimuli with emotional significance, their data "cannot speak to the issue of causality."[195] Thus, it is not clear what emotions are implicated, and no reason is given to presume that the results are attributable to racial animus as opposed to guilt, shame, or another emotion.

The neuropsychological research is in sharp tension with the implicit tone of moral condemnation inherent in the term "implicit prejudice." If we assume that spreading semantic activation and amygdala activity are beyond one's conscious control, can we hold others blameworthy for such factors as "bad" amygdala behavior? Note that Phelps and colleagues failed to find any relation between scores on explicit measures of prejudice and "bad" amygdala behavior.[196] If persons exhibit no explicit prejudice and their behavior is above psychometric reproach, but their amygdala fires suspiciously, what moral stance should be taken toward these individuals? We question whether they should be censured for manifesting the "residues" of a racist culture, particularly when their conscious attitudes indicate disgust or unhappiness with those aspects of American cultural history.

d. *Cultural Knowledge and Other Extrapersonal Associations, Not Personal Animus*

Soon after the IAT was introduced, researchers skeptical of the rush to label IAT scores evidence of implicit prejudice hypothesized that IAT results

---

enormous category mistake when we posit crude reductionist equivalences between neuronal activity and social-legal concepts such as moral responsibility, prejudice and racism. *See, e.g.*, MICHAEL S. GAZZANIGA, THE ETHICAL BRAIN 100 (2005) ("Responsibility is a human construct that exists only in the social world, where there is more than one person. It is a socially constructed rule that exists only in the context of human interaction. No pixel in a brain scan will ever be able to show culpability or nonculpability."); *id.* at 106 ("Even if [brain imaging] studies . . . eventually prove able to link specific brain activity to racist thoughts, it will remain difficult to prove that a person's racist thoughts necessarily lead to racist acts. The very suggestion is prejudicial and dangerous. This sort of one-to-one correspondence will prove wrongheaded in the long run.").

[193] Elizabeth A. Phelps et al., *Performance on Indirect Measures of Race Evaluation Predicts Amygdala Activation*, 12 J. COGNITIVE NEUROSCIENCE 729, 732–33 (2000) (finding this result in Experiment 1, but not Experiment 2).

[194] *Id.* at 734.

[195] *Id.*

[196] *Id.* at 732–33.

may be driven less by what respondents actually feel or believe and more by what they think other people—in the past or present—feel or believe (in short, cultural stereotypes to which respondents have been exposed, even in the context of public education campaigns designed to disabuse people of prejudice). Just as there is no good theoretical or methodological reason for supposing that the IAT is sensitive enough to distinguish sympathy from antipathy, there is no good reason for supposing it is sensitive enough to distinguish between automatically activated associations that, if called to people's attention, they would endorse and those associations that, if called to people's attention, they would categorically reject. Indeed, implicit-prejudice researchers make precisely this point when they declare that the associations needed to drive the IAT do not require conscious endorsement. It follows that mere knowledge of cultural stereotypes about minorities should be sufficient to cause people to manifest "prejudice" on the IAT.[197] Several studies support this hypothesis.[198] For instance, Judd, Blair, and Chapleau concluded that the activation of cultural stereotypes rather than activation of personal beliefs or prejudice better explains the anti-Black "shooter bias" found in the handgun identification studies referenced earlier.[199]

More generally, "the IAT may be influenced not only by information that is prevalent in our culture, but also by information that, although not the basis of one's personal evaluations and irrelevant to any privately made

---

[197] *See* Banaji et al., *supra* note 29, at 280 ("To speak of implicit attitudes as endorsed would be as nonsensical as speaking about a dog endorsing a bone."). Note that Wheeler and Petty, in their review of the effects of stereotype activation on behavior, concur that a stereotype can influence behavior even if one disagrees with the stereotype. *See* S. Christian Wheeler & Richard E. Petty, *The Effects of Stereotype Activation on Behavior: A Review of Possible Mechanisms*, 127 PSYCHOL. BULL. 797, 820–21 (2001).

[198] *See* Han et al., *supra* note 144, at 269 ("The two experiments reported here suggest that the traditional IAT is influenced both by one's personal associations and by extrapersonal associations—ones that are attitude-irrelevant but that are valenced and available in memory."); Karpinski & Hilton, *supra* note 131, at 786 ("[T]he results of these studies provide preliminary support for the environmental association model of the IAT. According to this model, the IAT taps the associations a person has been exposed to in his or her environment, not that individual's level of endorsement regarding the attitude object."); Michael A. Olson & Russell H. Fazio, *Reducing the Influence of Extrapersonal Associations on the Implicit Association Test: Personalizing the IAT*, 86 J. PERSONALITY & SOC. PSYCHOL. 653, 663 (2004) (replicating and extending work of Karpinski and Hilton and concluding that "[d]ata from the four experiments reported here suggest that the IAT has the potential to be contaminated by associations that although available in memory are irrelevant to one's evaluation of the attitude object").

[199] Charles M. Judd et al., *Automatic Stereotypes vs. Automatic Prejudice: Sorting Out the Possibilities in the Payne (2001) Weapon Paradigm*, 40 J. EXPERIMENTAL SOC. PSYCHOL. 75, 80 (2004); *see also supra* note 146 (describing prior handgun identification studies).

approach-avoidance decision, nevertheless is valenced and available in memory."[200] Accordingly, Han, Olson, and Fazio found that experimental subjects exposed to valenced information inconsistent with the subjects' own personal associations with an attitude object showed a reduced IAT effect (that is, the IAT picked up the extrapersonal association and not just personal associations with an attitude object).[201] And Olson and Fazio have shown in other studies that traditional versions of the IAT likely overstate the prevalence of anti-Black attitudes because the traditional IAT is a crude measure of associations that cannot distinguish between personal attitudes toward an object and extrapersonal associations with the same object.[202]

e. *Knowledge of Depressing Realities, Not Approval of Those Realities*

This next objection is easily misunderstood: it posits that implicit prejudice, as now conceived, labels perfectly rational reactions to existing socioeconomic conditions as prejudiced. It is easy to twist this argument so that it reads "prejudice or discrimination is rational." The actual argument is, however, more subtle: if one accepts the approach taken by IAT advocates to the definition of implicit prejudice, then one also accepts that it is reasonable to set one's threshold for labeling people prejudiced so low that virtually everyone—even rational observers of the social scene—qualifies as prejudiced. Curiously, this was one of the few points of convergence in a recent debate in *Psychological Inquiry* over the lessons that can be drawn from the IAT research program.[203] The leading psychological defenders of the IAT, Banaji, Greenwald, and Nosek, and their critics seemed to agree that the IAT makes it possible to be a Bayesian bigot (that is, simply following Bayesian principles for updating beliefs and computing probabilities in an inegalitarian society should qualify one as a bigot under the implicit-prejudice view)—although they disagreed over whether this constituted a strength or a *reductio ad absurdum* of the IAT research program.[204]

---

[200] Han et al., *supra* note 144, at 261.

[201] *See id.* at 267.

[202] Olson & Fazio, *supra* note 198, at 663 ("In Experiments 1 and 2, White participants appeared less prejudiced on the modified Black-White IAT than on the traditional IAT. This finding is consistent with the reasoning that when completing a traditional Black-White IAT, information about society's negative portrayal of Blacks facilitates the process of assigning Blacks and unpleasant items to the same response key, hence creating more prejudiced attitude estimates.").

[203] *See* Banaji et al., *supra* note 29, at 83–85; Tetlock & Arkes, *supra* note 29, at 319.

[204] *See* Banaji et al., *supra* note 29, at 285. Arkes and Tetlock were of the view that if the theoretical logic of the IAT requires that Jesse Jackson be classified as prejudiced

To understand how easy it is to qualify as a Bayesian bigot within the IAT methodology, it is necessary to grant only the following, borderline platitudinous, observations about American society. First, although there is sharp disagreement about why there is so much racial inequality, there is little disagreement about the raw facts: across the spectrum, African-Americans have, on average, fewer of the good things in life (high incomes and net worth, college educations, etc.) and more of the bad things in life (higher rates of imprisonment, violence, drug abuse, out of wedlock births, etc.). Second, although psychologists debate exactly how adept people are at assessing natural patterns of covariation,[205] no one has seriously argued that racial inequalities of the magnitude present in American society would likely go unnoticed by the majority of the American population, or fail to be encoded in the associative networks on which the population relies to navigate the social environment (judgment calls on which neighborhoods are safer—and does this covary with race?; which schools have higher test scores—and does this covary with race?; and so on). Third, given all that psychologists have learned about the dynamics of associative learning over the past century—from Pavlov's salivating dogs to Skinner's pecking pigeons to computer simulations of semantic memory—it is inevitable that a substantial network of negative associations will build up over a lifetime in a society in which there are large intergroup inequalities charged with negative emotional significance.

This argument implies that if we accept the theoretical logic of the IAT at face value, people are guaranteed to fail the test whenever they satisfy three starkly minimalist conditions: (a) they live in a society with intergroup inequalities; (b) they are reasonably attentive observers of covariations between group membership and important social facts (crime, school failure, poverty, etc); and (c) they attach evaluative significance to these factual inequalities, regardless of how they might explain them. This argument also implies that IAT researchers have overlooked the most obvious of all possible confounding variables within their associationist/learning-theory paradigm: knowledge of the social world. The more people know about the past and present history of American race relations, and about current patterns of inequality, the worse they should score on the IAT. It reveals much about the collective mindset of the IAT research community that no

---

against Blacks, then absurdity cannot be far off. *See* Tetlock & Arkes, *supra* note 29, at 319. Banaji et al. were of the view that changing times—scientific and societal—require changing definitions of prejudice and that there is no place for nostalgic clinging to old-fashioned definitions of prejudice. *See* Banaji et al., *supra* note 29, at 286–87.

[205] *See, e.g.*, Jonathan J. Koehler, *The Base Rate Fallacy Reconsidered: Descriptive, Normative, and Methodological Challenges*, 19 BEHAV. & BRAIN SCI. 1 (1996).

reported IAT study has yet implemented this fundamental control variable.[206]

The analytical distinctions in this subsection are of more than philosophical interest. Recognizing that real-world variables do not exist in statistical isolation from one another, the courts permit employers to base decisions on job-relevant attributes correlated with protected-category membership,[207] though such decision-making may expose the employer to disparate impact liability or disparate treatment liability if the correlated category is used as cover to discriminate on the basis of protected category membership.[208] Just as it would be bizarre to constrain employers to base

---

[206] McCauley and Stitt provide a measure for implementing this control variable which might be called an "explicit association test." *See* Clark R. McCauley & Christopher L. Stitt, *An Individual and Quantitative Measure of Stereotypes*, 36 J. PERSONALITY & SOC. PSYCHOL. 929, 933 (1978). These researchers examined the accuracy of stereotypes about African-Americans by comparing beliefs about what percentage of African-Americans and other Americans exhibit various population characteristics (e.g., high school graduate, unemployed, crime victim) with census data on these characteristics. *See id.* at 936. They found a high correlation between beliefs and data, indicating that the racial stereotypes were quite accurate. *See id.* at 937 ("Stereotyping on our seven characteristics is clearly veridical in terms of the direction of the difference perceived between black Americans and all Americans .... [T]he diagnostic ratios do not appear more extreme than the criterion ratios."). *But see* MARTIN GILENS, WHY AMERICANS HATE WELFARE 68 (1999) (presenting survey evidence in which Americans overestimated the poverty rate among African-Americans); Ted Chiricos et al., *Racial Typification of Crime and Support for Punitive Measures*, 42 CRIMINOLOGY 359, 369–70 ("Estimates substantially exaggerate black involvement in violent crime, slightly exaggerate black involvement in burglary and underestimate black involvement in robbery."). Given the psycho-logic of the IAT, subjects in the McCauley and Stitt study should have exhibited the IAT effect, regardless of their animus toward African-Americans, simply because they recognized that, for whatever reasons, African-Americans suffer from social problems at a greater rate than other groups in American society.

[207] In *Hazen Paper*, the Court ruled that evidence an employer acted on the basis of an age-correlated variable did not prove intentional age discrimination, at least where age and the other variable are "analytically distinct." Hazen Paper Co. v. Biggins, 507 U.S. 604, 611 (1993) ("Because age and years of service are analytically distinct, an employer can take account of one while ignoring the other, and thus it is incorrect to say that a decision based on years of service is necessarily 'age based.'"); *see also* Hernandez v. New York, 500 U.S. 352 (1991) (permitting jury strikes on the basis of Spanish-English bilingualism despite their adverse impact on persons of Hispanic national origin).

[208] Reliance on a variable correlated with protected-category membership may give rise to a prima facie case of disparate impact, but the employer may avoid liability by showing that reliance on this correlated variable is a business necessity or has been vetted as part of a professionally developed test process. Disparate impact claims are viable under Title VII, the ADEA, and the ADA, and "it seems likely that disparate impact remains available only to racial minorities and women, with the possible exception that

their decisions solely on variables that have zero correlations with membership in protected groups, it would be bizarre to expect people to fail to notice real-world statistical relationships involving protected categories—and to expect them not to form mental models of the world (associative networks) that reflect those relationships. At best, the IAT demonstrates that these mental models of correlated variables exist for some people—it cannot disentangle the causal role that either variable, or the correlation itself, plays in judgment and choice, and it does not take into account competing mental models that can be readily primed in other contexts.[209]

The importance of this distinction can be brought home with a thought experiment. It is highly likely that the IAT researchers who report "failing" the IAT, including the test's developers, Anthony Greenwald and Mahzarin Banaji,[210] do so because they recognize the plight of minorities and sympathize with those groups—indeed, their research is often motivated by a desire to rectify inequalities.[211] Yet, by the logic of the IAT, these researchers are implicit bigots on par with children reared in prejudiced households and taught to hold mean-spirited beliefs about minorities and to act out these prejudices. But do we really believe that these researchers and others who "fail" the IAT because they sympathize with minorities—the

---

whites and males can utilize the theory when they, too, are minorities in a particular workplace." Charles A. Sullivan, *The World Turned Upside Down?: Disparate Impact Claims by White Males*, NW. U. L. REV. 1543, 1543 (2004). In addition, conscious use of, say, income as a proxy to discriminate on racial or ethnic grounds will give rise to a disparate treatment claim. *See Hazen Paper Co.*, 507 U.S. at 611 ("The employer cannot rely on age as a proxy for an employee's remaining characteristics, such as productivity, but must instead focus on those factors directly.").

209 We discuss work on the power of realistic work-setting cues to prime very different, sometimes pro-Black, sets of associations in the section on external validity. *See infra* Part III.D.

210 *See* Shankar Vedantum, *See No Bias*, WASH. POST, Jan. 23, 2005, at W12 (relating story about failed IATs by Professors Banaji and Greenwald, developers of the IAT); *see also* AYRES, *supra* note 103, at 427 ("I'm saddened to report that after taking the implicit attitude test several times, I have consistently been rated as having a 'moderate' to 'strong' automatic preference for whites relative to blacks . . . . I find my IAT results jarring not because I am surprised that growing up in this society almost necessarily racialized my perceptions, but because I was mistakenly confident that my nondiscriminatory ego could stamp out any unconscious discriminatory predispositions.").

211 *See* Dasgupta, *supra* note 36, at 144 ("The discrepancy between increasingly tolerant self-reported attitudes in the face of enduring and glaring disparities in people's lived experience prompted some social psychologists to urge the development of alternative, less obtrusive, measures of attitudes and behavior that do not rely so heavily on people's willingness and ability to accurately self-report their thoughts and actions, especially with regard to socially sensitive issues like prejudice and stereotypes." (citations omitted)).

same persons who often endorse affirmative action programs, as Kang and Banaji do,[212] and who actively police themselves for discrimination—are likely to make employment decisions that disadvantage minorities relative to majority groups? Even in situations involving subjectivity, which antidiscrimination scholars and IAT proponents most worry about,[213] it is more probable that implicit associations entwined with sympathy for minorities will work for, not against, minorities in the workplace.

f. *Cognitive Dexterity, Not Antipathy*

We close this section by adding one more explanatory contender to the mix: the possibility that the IAT will prove a better measure of cognitive flexibility than of prejudice.

Recall that one's score on the IAT is a difference score: how long it takes to respond on "incompatible" trials (e.g., "Black" names with positive adjectives) minus how long it takes to respond on "compatible" ones (e.g., "White" names with positive adjectives). If incompatible trials take longer than compatible trials, then IAT defenders treat this difference as evidence of implicit prejudice or stereotypes at work. A curious implication of this procedure is that rapid responding on compatible trials contributes as much to one's score as sluggish responding on incompatible trials. All trials—being fast on the designated response keys for White + positive and Black + negative in "compatible" trials, and being slow on the designated keys for White + negative and Black + positive in "incompatible" trials—receive identical weights. IAT researchers recognize some of the dangers created by this procedure. They recommend data-transformation procedures to minimize the influence of individual differences in cognitive processing speed. They argue quite rightly that, by standardizing each individual's scores around his or her own personal average reaction time, researchers can reduce the risk that low scores on the IAT are confounded with individual differences in ability to respond rapidly to any set of stimuli that might flash across a computer screen.[214]

---

[212] *See, e.g.*, Kang & Banaji, *supra* note 3, at 1098–99 (arguing for tie-breaking in favor of bias targets).

[213] *See, e.g.*, Hart, *supra* note 151, at 742 ("A decisionmaking process where the subjective judgments of the selecting officials are the primary criteria is particularly at risk for [unconscious] discrimination."); Kang & Banaji, *supra* note 3, at 1074 (warning against subjectivity in employment interviews).

[214] However, Blanton and colleagues have recently shown that the use of a difference score "can only remove processing speed confounds [from the IAT] under restrictive conditions." Blanton et al., *supra* note 173, at 208; *see also id.* at 209–10 (appendix discussing strategies for controlling general processing speed confounds).

There is, however, one deeply threatening confound that no amount of data transformation can eliminate: the possibility that a personality dimension other than prejudicial tendencies influences performance on all speeded-classification tasks that require flexibly switching response-preparedness (be it responding left-right or up-down to different stimuli, whatever their content). Research on cognitive styles strongly suggests that such individual differences exist and arise from variation in mental flexibility, fluid intelligence, and hand-eye coordination rather than from attitudinal variation. And psychologists have been on this trail for a long time. In 1965, Baeumler advanced the notion of "interference inclination" to explain individual differences in the capacity to respond rapidly on tasks that require inhibiting competing response inclinations at different moments.[215] There is also direct evidence from recent work. Mierke and Klauer experimentally created associations between target stimuli (blue versus red) and attribute stimuli (large versus small) and then programmed an IAT to measure these arbitrary geometric associations.[216] They found that scores on this "geometric association" IAT correlated with individual differences on the original "flower-insect" IAT that defined the procedural template for IAT work on implicit prejudice.[217] Evidence of a systematic individual difference across types of IATs strongly suggests the operation of a general personality factor that, independently of prejudice, shapes performance on all speeded binary classification tasks.

5. *Mystery Metric: Reaction Time Does Not Possess Intrinsic Social Meaning Even if We Assume that Reaction Times Reflect Association Strength*

Even if we discard all of the foregoing grounds for concern and accept that the IAT is a valid measure of implicit prejudice, a vexing and thus far unsolved psychometric problem remains: namely, that the IAT employs an arbitrary metric and IAT scores, by themselves, say nothing about a particular individual's propensity to discriminate. Only by linking IAT scores

---

[215] *See* Fiedler et al., *supra* note 188, at 17, 24 (citing G. Baeumler, *Interferenz und Intelligenz*, 8 PSYCHOLOGISCHE BEITRÄGE 596 (1965)).

[216] *See* Jan Mierke & Karl Christoph Klauer, *Method-Specific Variance in the Implicit Association Test*, 85 J. PERSONALITY & SOC. PSYCHOL. 1180, 1189–90 (2003).

[217] In the flower-insect IAT, participants tend to respond more quickly when flowers are paired with pleasant-meaning words and insects are paired with unpleasant-meaning words. The flower-insect IAT produced the first reported results from the IAT as a supposed implicit measure of attitudes, in that case of attitudes towards flowers and insects. *See* Greenwald et al., *supra* note 5, at 1466–70.

to reliable patterns of observable behaviors can specific IAT scores acquire diagnostic meaning.

As pointed out by Brendl and his colleagues, the IAT provides

> [A]t best a relative measure of one target set against another. However, in contradiction to this constraint of relativity, the results of the IAT are often interpreted as reflecting an implicit prejudice for one group over another. The problem with this interpretation is that . . . prejudice connotes a negative attitude toward a group.[218]

Assume that a subject taking the IAT responds more slowly when the name *Ebony* requires responding to the *Good* key than when the name *Betsy* requires responding to the *Good* key.[219] Also grant the assumption that relative reaction time reflects associations between these names and positive adjectives and these associations in turn connote an attitude toward groups for which these names are more common. It remains possible that the subject holds an implicit positive attitude toward African-Americans, albeit not as positive as toward Whites. A relative difference in reaction time between two target sets does not necessarily imply hostility or prejudice toward either group.

But the psychometric problems run still deeper. Presently, no scale or function exists for translating different reaction times into differences in direction, intensity, or strength of attitudes across persons. It is possible that, for many samples and contexts, a reaction-time differential of plus or minus 200 milliseconds around the IAT zero point of perfect color-blindness has no attitudinal or behavioral implications whatsoever. By contrast, for other samples and contexts, differentials of this size may carry great significance. Unfortunately for those eager to extract strong prescriptive conclusions from this body of work, implicit prejudice researchers have yet to theoretically specify, less still empirically quantify, the functional relationships between reaction times and attitudinal positions. Stated simply, we have no idea whether the same reaction-time differences across persons, samples or contexts have the same attitudinal implications. Ignorance on this scale is disconcerting when law review proponents of the IAT use the test to justify sweeping assertions about the pervasiveness and potency of implicit biases.

Absent full disclosure of these limitations, the IAT will appear much more scientific to outside observers than it actually is. Although reaction time looks like an impressive ratio-scale metric, with all the objective

---

[218] Brendl et al., *supra* note 71, at 771.

[219] This example is taken from the original race IAT, which utilized names that would supposedly trigger different racial associations. *See supra* Part II.B.

properties of temporal measurements in physical science, it is an utterly arbitrary metric for its intended purpose in the IAT: the assessment of the not-directly-accessible psychological construct of implicit prejudice. For instance,

> [I]t is not known where a given [reaction time] locates an individual on the underlying psychological dimension or how a one-unit change [in observed reaction time] reflects the magnitude of change on the underlying dimension . . . . When a metric is arbitrary, the function describing this relationship and the parameter values of that function are unknown.[220]

With an arbitrary metric, researchers must meticulously map scores on the metric to observable behaviors or external events to lend meaning to the otherwise arbitrary metric. "In the case of the race IAT, this means that [the IAT's reaction-time] metric becomes meaningful to the extent that one knows just how much relative implicit preference for Whites versus Blacks is implied by any given IAT score."[221]

Instead of undertaking this mapping task, IAT researchers have simply assumed that identical reaction times across compatible and incompatible trials reflect no preference for majority or minority groups and differential reaction times reflect relative preferences for the groups. "Although studies have investigated the predictive validity of the race IAT with regard to racial attitudes and prejudicial behaviors, no published study has shown that the zero-point used to diagnose 'attitudinal preferences' is the true dividing line between preference for Blacks versus Whites."[222] In view of our earlier demonstration that systematic factors other than attitudinal preference demonstrably influence reaction times on the IAT, and that reaction time scores contain a great deal of random and systematic measurement error, it is inappropriate to treat reaction times as unproblematic measures of relative liking of majority and minority groups.[223]

---

[220] Hart Blanton & James Jaccard, *Arbitrary Metrics in Psychology*, 61 AM. PSYCHOL. 27, 28 (2006).

[221] *Id.* at 13.

[222] *Id.* at 17.

[223] *See id.* at 16; *see also* Sam G. McFarland & Zachary Crouch, *A Cognitive Skill Confound on the Implicit Association Test*, 20 SOC. COGNITION 483, 503–04 (2002) ("As traditionally administered, with many exemplars in each category, IAT scores are confounded by a respondent's skill in responding on control IATs . . . . The magnitude of the confound is substantial . . . ."); *id.* at 504 ("The implications of this confound are straightforward: Participants with little difference in their response speeds to incongruent versus congruent pairings on control IATs are biased toward lower prejudical IAT scores; those with larger differences are biased toward higher IAT-assessed prejudice."). The use of standardized difference scores to decide relative preference on the IAT further

complicates matters. Persons whose reaction times vary little in response to majority and minority group targets could receive an IAT effect size equal to or larger than a person with widely varying response times because the first subject's raw scores are divided by a much smaller standard deviation. *See* Blanton & Jaccard, *supra* note 220, at 35.

Building on Blanton and Jaccard's demonstration of the need to remove confounding factors from IAT scores, *see id.* at 40–41, we illustrate below the many assumptions that must hold to give the zero-prejudice meaning to the zero point on the IAT scale. The first and second equations represent, respectively, the factors influencing reaction times on incompatible (IRL) and compatible (CRL) trials, where GPS represents general processing speed of respondents, A_W represents attitudes toward Whites, B_W represents attitudes toward Blacks, F represents familiarity, EE represents egalitarian empathy, CK represents cultural knowledge (especially of stereotypes), PK represents political knowledge (especially of depressing socio-economic facts), IT represents identity threat (fear of being labeled a bigot), the ε's represent random measurement error, the β's represent slopes, and the α's represent intercepts:

$$IRL = \alpha_I + \beta_1 GPS + \beta_2 A_W + \beta_3 A_B + \beta_4 F + \beta_5 EE + \beta_6 CK + \beta_7 PK + \beta_8 IT + \varepsilon_I$$

$$CRL = \alpha_c + \beta_9 G + \beta_{10} A_W + \beta_{11} A_B + \beta_{12} F + \beta_{13} EE + \beta_{14} CK + \beta_{15} PK + \beta_{16} IT + \varepsilon_I$$

It follows that the difference in reaction times between incompatible and compatible reaction times, which is the most widely used index of implicit prejudice, must be a function of the following:

$$IRL - CRL = (\alpha_C - \alpha_I) + (\beta_1 - \beta_9)GPS + (\beta_2 - \beta_{10})A_W + (\beta_3 - \beta_{11})A_B + (\beta_4 - \beta_{12})F + (\beta_5 - \beta_{13})EE + (\beta_6 - \beta_{14})CK + (\beta_7 - \beta_{15})PK + (\beta_8 - \beta_{16})IT + (\varepsilon_C - \varepsilon_I)$$

Therefore, researchers who assume that the zero point on the IAT scale indicates zero prejudice must also assume that the following difficult-to-satisfy psychometric conditions have been satisfied:

1) $\alpha_C = \alpha_I$
2) $\varepsilon_C = \varepsilon_I$
3) $\beta_1 = \beta_9$
4) $\beta_4 = \beta_{12}$
5) $\beta_5 = \beta_{13}$
6) $\beta_6 = \beta_{14}$
7) $\beta_7 = \beta_{15}$
8) $\beta_8 = \beta_{16}$
9) $\beta_{10} > \beta_2$
10) $\beta_{11} < \beta_3$
11) $|\beta_{10} - \beta_2| \geq |\beta_{11} - \beta_3|$

This list actually underestimates the number of assumptions underpinning the "zero point = no prejudice" interpretation of the IAT. If we allow for interactions among known causal factors, the list easily triples in length. Because of the considerable difficulty in testing the influence of all of these variables and their interactions in a unified study, Blanton and Jaccard sensibly argue for the alternative strategy of mapping IAT scores

6. *Recapitulation of Construct Validity Arguments: What Still Stands?*

What proportion of the variance can be plausibly attributed to unconscious prejudice after we subtract the variance properly assignable to the six counter-interpretations considered here? The direct answer is: no one knows. And that, in itself, should suffice to put the brakes on legal applications of this line of work. For our part, we feel it is safe to propose the reputational bet that the answer will prove considerably closer to 0% than to the 100% "pure-prejudice" position taken by Banaji and her colleagues. This reputational bet strikes us as safe in view of the fact that reliability studies reveal half or more of the variance in the IAT to be error variance and that construct-validity studies reveal the need to apportion significant amounts of the remaining variance to each of the six alternatives advanced here, with high likelihood that other viable alternative explanations will yet be proposed.

B. *Internal Validity: Does Implicit Prejudice Cause Discriminatory Behavior?*

When implicit prejudice researchers find a predicted correlation between their hypothesized causal variable (implicit prejudice) and one of its hypothesized effects (say, increased eye blinking), they typically argue that this predictive success increases both the plausibility of the claim that their measure taps into implicit prejudice and of the claim that implicit prejudice causally contributes to the hypothesized effect.[224] This chain of logical inference breaks down, however, insofar as the research design lacks internal validity, or the power to rule out alternative causal interpretations of the patterns of covariation in the data. Implicit prejudice research is plagued by internal validity problems for two basic reasons.

First, and most obvious, correlation does not prove causation. One set of possibilities, raised repeatedly in the construct validity section, is that there is so much unresolved controversy about what implicit measures of prejudice assess (familiarity, guilt, embarrassment, knowledge of stereotypes, knowledge of the environment) that it is rash to assume that implicit prejudice is the best explanation for correlations between measures of "implicit prejudice" and criterion variables. However, even if we had platonically pure measures of implicit prejudice, it would still be necessary to show that correlations between such measures and criterion variables hold up after controlling for the influence of alternative psychological constructs.

---

onto observable events that should differ in respect to the zero point. *See* Blanton & Jaccard, *supra* note 130, at 68, 70.

[224] *See supra* Part III.A.3.

The list of overlapping constructs is a long one, but the most glaring omission in many studies is the failure to show that implicit measures of prejudice have predictive power for criterion variables after controlling for old-fashioned explicit measures of prejudice.[225] One could argue that this methodological omission is relatively harmless because implicit and explicit measures are dissociated, as some researchers have reported.[226] But, as noted earlier, other researchers argue that implicit and explicit measures have considerable statistical overlap and that this constitutes evidence for the convergent validity of the measures.[227] The net result is that implicit prejudice researchers are caught between a rock and a hard place. If they argue for the full-dissociation thesis, they undercut the convergent-validity argument. But if they make strong convergent-validity claims, they raise the question of whether their effects would hold up after controlling for the influence of moderately to highly correlated explicit measures. It is odd, after all, to claim to have discovered a new causal source of racial discrimination without statistically controlling for the possibility that the effects can be explained by traditional sources of racial discrimination.

Second, internal validity problems are daunting because the criterion variables typically chosen in predictive validity research are open to multiple interpretations.[228] For example, the criterion variables in many studies involve measures of the quality of an interaction between subjects and members of minority versus majority groups.[229] When a subject exhibits

---

[225] And in one of those rare cases where the researchers did check, the correlation between IAT scores and expressed preferences for interactions with a White or Black partner decreased slightly and passed from significance to marginal significance after controlling for explicit racial attitudes. *See* Ashburn-Nardo et al., *supra* note 156, at 76.

[226] *See, e.g.*, Rudman, *supra* note 76, at 132 ("[A]bsence of consistently strong [Implicit Explicit] convergence underscores the discriminant validity of response latency techniques . . . ."); Rudman & Glick, *supra* note 154, at 755 ("[T]he relationships among the IAT and the explicit measures were negligible for both men and women, supporting their discriminant validity." (footnote omitted)).

[227] *See supra* Part III.A.2.

[228] For an overview of the predictive validity studies, see *supra* Part III.A.3.

[229] For example, Dovidio and colleagues found that the number of eye blinks and amount of eye contact toward an African-American versus a White person correlated with levels of implicit prejudice. *See* John F. Dovidio et al., *On the Nature of Prejudice: Automatic and Controlled Processes*, 33 J. EXPERIMENTAL SOC. PSYCHOL. 510 (1997) (Experiment 3). Similarly, McConnell and Leibold found that pro-White bias on the IAT related to differences in some nonverbal behaviors during interactions with a Black versus a White Experimenter (significant differences were found with respect to speech errors, smiling, and speaking time, but not for seating distance, fidgeting, or expressiveness). *See* McConnell & Leibold, *supra* note 85, at 439–40. Other predictive validity studies have used dependent measures such as judgments of responsibility and

greater interpersonal anxiety or greater distance in the presence of a minority group member and this subject also scores as implicitly prejudiced on the IAT, one explanation is that this behavior represents unconscious prejudice leaking out into interpersonal relations.[230] Until alternative explanations are eliminated, however, the internal validity of such arguments is suspect.

Obvious alternative explanations for the effects in these studies await testing. Certainly bigotry may cause Whites to sit further from or avoid eye contact with African-Americans. But so too may shame. Indeed, Keltner has found that a downcast gaze, halting speech, verbal silence, and slumped posture are characteristics of shame.[231] Thus, a White person who is genuinely ashamed of society's treatment of African-Americans, or perhaps fears that African-Americans are understandably wary of Whites, might well be scored as prejudiced by raters in validation studies that probe links between implicit prejudice and nonverbal behavior. Yet a person who is ashamed of Whites' treatments of African-Americans is not likely to be a bigot; rather, the opposite is likely.

Another possibility involves social awkwardness stemming from lack of experience with members of other ethnic-racial groups. Consistent with this explanation, Keltner and Buswell found that a downcast gaze is also characteristic of embarrassment,[232] and Asendorpf collected evidence

---

emotional reactions to racially-charged stimuli. For example, Fazio and colleagues reported a significant correlation between one's results on an affective-priming task and one's assignment of responsibility for the 1992 Los Angeles riots primarily to African Americans. *See* Fazio & Olson, *supra* note 5, at 305. Using the affective-priming procedure, Fazio and Hilden were able to predict emotional reactions to a public service advertisement that led viewers to draw an unwarranted and prejudiced conclusion. *See* Russell H. Fazio & Laura E. Hilden, *Emotional Reactions to a Seemingly Prejudiced Response: The Role of Automatically Activated Racial Attitudes and Motivation to Control Prejudiced Reactions*, 27 PERSONALITY & SOC. PSYCHOL. BULL. 538 (2001). And Fazio and Dunton reported a relation between racial attitudes detected by the affective-priming procedure and the extent to which racial characteristics—as opposed to occupational or gender-related ones—were used to assess the similarity of photographs. *See* Russell H. Fazio & Bridget C. Dunton, *Categorization by Race: The Impact of Automatic and Controlled Components of Racial Prejudice*, 33 J. EXPERIMENTAL SOC. PSYCHOL. 451 (1997).

[230] Kang, *supra* note 16, at 1524 ("These nonverbal behaviors that leak out from our implicit bias influence the quality of our social interactions." (footnote omitted)).

[231] *See* Dacher Keltner & Brian N. Buswell, *Evidence for the Distinctiveness of Embarrassment, Shame, and Guilt: A Study of Recalled Antecedents and Facial Expression of Emotion*, 10 COGNITION & EMOTION 155 (1996); Dacher Keltner & L. Harker, *The Forms and Functions of the Nonverbal Signal of Shame*, *in* SHAME: INTERPERSONAL BEHAVIOR, PSYCHOPATHOLOGY, AND CULTURE 78 (P. Gilbert & B. Andrews eds., 1998).

[232] Dacher Keltner & Brian N. Buswell, *Embarrassment: Its Distinct Form and Appeasement Functions*, 122 PSYCHOL. BULL. 250 (1997).

showing that speech disturbances also characterize embarrassment.[233] Given the segregated nature of many American high schools, a White undergraduate student being interviewed by an African-American experimenter might find that situation to be an unfamiliar one that fosters anxiety and embarrassment. Of course, a person experiencing such emotions and displaying awkward nonverbal behaviors is not necessarily prejudiced.

Defenders of the implicit prejudice research program could argue that, although the nonverbal behaviors in some studies might reasonably be attributed to shame, embarrassment, or some other emotion, prejudice is the most parsimonious explanation because it is a possible cause of the targeted nonverbal behavior in all of these studies. However, given that the overwhelming majority of White undergraduates score quite low on explicit measures of prejudice,[234] it is just as parsimonious to suppose that guilt and shame over their race's past treatment of African-Americans would be aroused in these situations. Neither the IAT nor the affective-priming methods can answer these fine-grained questions about which motive underlies which nonverbal tic or twitch. Yet some researchers insist that their results specifically tap implicit prejudice, rather than guilt, nervousness, or other automatically activated reactions.[235]

It is unclear to us how data such as differential eye gaze duration or eye blink frequency can be confidently attributed to implicit prejudice, given that these nonverbal behaviors have long been known to be characteristic of a host of other affective states. Unconscious prejudice is neither the only nor the most plausible explanation for such findings.

C. *Statistical Conclusion Validity: Is Implicit Prejudice Really Pervasive and Distinct from Explicit Prejudice?*

Implicit-prejudice research has become a sociological phenomenon because the researchers claim to have developed mindreading tools that reveal a great deal more prejudice in modern American society than has been registered by conventional opinion surveys.[236] It is doubtful the implicit

---

[233] Jens B. Asendorpf, *The Expression of Shyness and Embarrassment*, *in* SHYNESS AND EMBARRASSMENT: PERSPECTIVES FROM SOCIAL PSYCHOLOGY 87 (W. R. Crozier ed., 1990).

[234] *See, e.g.*, Monteith et al., *supra* note 2.

[235] *See, e.g.*, Dovidio et al., *supra* note 229.

[236] *See, e.g.*, BBC News, *Ten Minute Test Could Spot Killers*, http://news.bbc.co.uk/1/hi/health/2943160.stm (last visited Oct. 14, 2005) ("The 10 minute test is based on the Implicit Association Test, developed in the United States, and used to reveal people's deepest thoughts and feelings."); Vedantum, *supra* note 210, at W13 (This *Washington Post* reporter writes that "results of the millions of [IAT] tests

that have been taken anonymously on the Harvard web site and other sites hint at the potential impact of the research. Analyses of tens of thousands of tests found 88 percent of white people had a pro-white or anti-black implicit bias; nearly 83 percent of heterosexuals showed implicit biases for straight people over gays and lesbians; and more than two-thirds of non-Arab, non-Muslim volunteers displayed implicit biases against Arab Muslims."); *id.* ("The [IAT] bias tests . . . have arguably revolutionized the study of prejudice. In their simplicity, the tests have raised provocative questions about this nation's ideal of a meritocracy and the nature of America's red state/blue state political divide.").

Such remarks could—and should—be written off as journalistic hyperbole. But what should we make of Professor Banaji's claim that the IAT is to psychological research on prejudice what Galileo's telescope was to the Copernican Revolution? *See* Jill D. Kester, *A Revolution in Social Psychology*, APS OBSERVER ONLINE (July/Aug. 2001), http://www.psychologicalscience.org/observer/0701/family.html. She additionally claims that the IAT ushers in a scientific revolution that will be harder to accept than the Copernican or Darwinian revolutions:

> [It] will challenge our beliefs about the very nature of our own minds . . . it is not merely about the place of our planet amongst other planets, it's not merely about our place in the larger set of other species, it's about the core issue of our competence, it's about our goodness, our ability to be moral, and to have control over our thoughts and feelings, about the most important object in our universe, other humans.

Kester, *supra*; *see also* Banaji et al., *supra* note 29, at 281 ("Moving from Newtonian physics to quantum mechanics required large shifts in assumptions, technology, and understanding. There is no reason to assume that the smaller steps in any science that moves away from the familiar and comfortable (here, the view of prejudice as only conscious) is any different.").

We would caution that greater historical distance is needed for distinguishing genuine scientific revolutions, which are rare, from the fads and illusions of progress that regularly roil the behavioral sciences. We would also note that the IAT-telescope analogy reveals a lack of appreciation for arguably the most profound difference between the physical and social sciences: the far looser conceptual connections between measurement and target constructs in the social sciences. Labeling the IAT a measure of "implicit" "prejudice" is, twice over, an act of scientific over-claiming. First, as a leading implicit-memory researcher, Larry Jacoby, has pointed out, using the adjective "unconscious" or "implicit" to refer to target constructs requires making a "process-pure" assumption (that the construct validation problem has been solved and that the implicit measure taps implicit processes, the unobservable target construct). *See* Larry L. Jacoby, *Dissociating Automatic and Consciously Controlled Effects of Study/Test Compatibility*, 35 J. MEMORY & LANGUAGE 32, 34–35 (1996); *see also* Conrey et al., *supra* note 163, at 470 ("The more general point is that no task is 'process pure.' It is technically impossible that any task that requires observable responses depends entirely on automatic processes and not at all on controlled processes."). It is more accurate to characterize the IAT as an implicit measure of a difficult-to-disentangle mixture of processes or states, some of which may be implicit, a further subset of which may have an evaluative component, and a further subset of which may tap into affect that observers of a certain political orientation might characterize as prejudicial. Further, as De Houwer recently demonstrated, the assumption that the IAT measures only automatic associations, without

prejudice research would have attracted the massive media attention it has if researchers had been more circumspect in the labels they attached to their results and in the scope claimed for implicit prejudice. The *New York Times*, *Boston Globe*, *NBC Dateline*, *CNN* and other media outlets would certainly have shown less interest if social psychologists had offered a more theoretically modest and technically accurate designation of the IAT: a measure of automatically activated affect that may be intertwined with a wide range of emotionally charged appraisals of target groups (from sympathy to contempt) and that predicts behavioral criterion variables of uncertain meaning under yet-to-be-determined boundary conditions.[237]

Additional concerns arise about these far-reaching claims when we evaluate implicit prejudice research in terms of our third category of validity, statistical conclusion validity. The data analyses used by implicit prejudice researchers suffer from three recurring flaws: (a) over-reliance on low to moderate correlation coefficients for making claims about the pervasiveness of prejudice and about the propensities of IAT test-takers to discriminate; (b) inattention to the potential role of outliers in biasing correlation coefficients and in rendering correlations even more misleading as estimates of population-wide propensities to discriminate; and (c) inattention to stimuli-

---

the influence of conscious propositional knowledge, is not correct. *See* Jan De Houwer, *Using the Implicit Association Test Does Not Rule Out An Impact of Conscious Propositional Knowledge on Evaluative Conditioning*, 37 LEARNING & MOTIVATION 176, 186 (2006) ("The present results make clear, however, that one cannot simply assume that implicit, reaction time based measures are impervious to the effects of conscious propositional knowledge.").

Second, as noted earlier, prejudice is not a natural phenomenon, such as light, that can be analyzed objectively into its components. Prejudice is, in part, a political construct that takes on radically different meanings at different times and places. Unlike cognitive neuroscience constructs such as amygdala activation, prejudice is not a purely intra-psychic construct that unfolds inside the bony confines of our craniums; it is the product of the broader battle of ideas, and struggle for power, in society as a whole. The telescope analogy stands as another warning to outsiders that some IAT researchers are using science for the "honorific" purpose of advancing a political agenda.

[237] *See* Implicit Association Test, In the Media, https://implicit.harvard.edu/implicit/canada/inthemedia.html (last visited Sept. 16, 2005). In addition to these media reports, it has become common for websites aimed at educating the public on intergroup relations to link to the IAT and information about research findings employing the IAT. *See, e.g.*, Bowling Green State University, Office of Equity & Diversity, http://www.bgsu.edu/offices/oed/page7695.html (last visited Sept. 21, 2005); Seattle Pacific University, Intercultural Affairs, Ethnic Minority Links, http://www.spu.edu/depts/intercultural/multiethnic/ethnicminority.asp (last visited Sept. 21, 2005); Tolerance.org, Dig Deeper: Test Yourself for Hidden Bias, http://www.tolerance.org/hidden_bias/ (last visited Sept. 21, 2005); Understanding Prejudice, Exercised and Demonstrations: Implicit Association Test, http://www.understandingprejudice.org/iat/ (last visited Sept. 21, 2005).

specific processes that limit the generality of statistical conclusions.

1. *Correlation Coefficients (of the Magnitude Observed in IAT Research) Shed Limited Light on the Relative Risks of Classification Errors*

A specious but seductive syllogism resides at the heart of the implicit prejudice argument that legal scholars wish to import into American law. The major premise of that syllogism is that the vast majority of the population harbors the morally corrosive psychological construct "implicit prejudice." We know this to be true because 80% or more of the population often scores as prejudiced according to the IAT.[238] The minor premise is that implicit measures of prejudice tend to have low but positive correlations with judgments or acts that could be construed as prejudicial toward African-Americans or other protected groups. We know this to be true from a recent meta-analysis which shows that IAT-criterion variable correlations hover in the vicinity of .25.[239] Accordingly, we are justified in concluding that the vast majority of the population will exhibit these same prejudicial tendencies.[240]

The conclusion does not follow from the premises: quite the opposite, the fact that IAT scores have low positive correlations with behavior expansively defined as discriminatory guarantees that many individuals labeled prejudiced by the IAT *will not* exhibit the behavior in question.[241] Indeed, we can estimate how common false accusations of prejudice will be if we make plausible simplifying assumptions about three parameters: (a) how often people "fail" the IAT by showing an implicit bias toward some group (let's

---

[238] *See supra* notes 78–80 and accompanying text.

[239] *See supra* note 96 and accompanying text.

[240] For instance, Richardson & Pittinsky write that:

> After several years and experiment modifications over time, [implicit measurement] methodologies have been accepted as both valid and reliable. With over 75,000 interpretable results, 75% of White participants and 42% of Black participants showed pro-White/anti-Black preference. These findings shatter the Court's understanding of prejudice and discrimination as being the result of the malicious intentions of a minority. They suggest that discriminatory attitudes are ubiquitous, often operating without the conscious awareness of the individuals harboring them.

Margaret Richardson & Todd L. Pittinsky, *The Mistaken Assumption of Intentionality in Equal Protection Law: Psychological Science and the Interpretation of the Fourteenth Amendment* 31 (KSG Faculty Research Working Paper Series RWPO5-011, 2005).

[241] For information on the correlation between IAT scores and discriminatory behavior, see *supra* Part III.A.3.

say, drawing on the published literature, the "failure" base rate ranges between 70% and 90%); (b) the hit rate of the test, or the probability that if someone is truly prejudiced that person will "fail" the IAT (let's stipulate that these hit rates range between 70% and 90%, which is generous given the IAT's low test-retest reliability and the number of confounding variables distorting IAT scores); and (c) the base rate of "true" prejudice in the population (let's say, drawing on explicit measures used in survey research, that these true-prejudice base rates range from 10% to 30%).

Using elementary Bayesian analysis, we find that false-accusation rates range between approximately 60% (when the failure rate of the IAT is 70%, the base rate of true-prejudice is 30%, and the IAT's hit rate is 90%) and 90% (when the failure rate of the IAT is 90%, the base rate of true prejudice is 10% and the IAT's hit rate is 70%).[242] Whether one judges such false-accusation rates to be excessive is ultimately a matter of political values, not scientific fact. But these rates should be sobering.

The most tempting defense for IAT proponents is to point out that much hinges on the extremity of the assumptions we choose to make about the base rate for the hypothetical construct of true prejudice in the population. IAT researchers, however, flirt with tautology if they argue that they know the base rate of true prejudice must be much higher than posited in the earlier estimates, as high as in the 70% to 90% range, because, after all, roughly those percentages of people fail the IAT. Unfortunately for IAT proponents, that counter-argument has force only if we grant IAT researchers the pivotal assumption that the IAT is a pure measure of prejudice, which is the central point in contention.

The published data on IAT-criterion variable correlations point to a possible path out of this morass. We can mathematically deduce the correlations that must hold between the IAT and hypothetical criterion variables capturing true prejudice when we make varying assumptions about the p(prejudice | IAT failure), the p(IAT failure), and the base rate of true prejudice, p(prejudice).[243] The results tell us that the population base rate of

---

[242] The estimates of 61% to 92% are derived from the most basic definitions of probability, as follows: p(not prejudiced |failure on the IAT) = 1 – p(prejudiced | failure on the IAT) = 1 – (p(failure on IAT | true prejudice) * p(true prejudice)/p(failure on the IAT)). Bayes' theorem provides the proofs of these results. *See* Richard Price, *An Essay Towards Solving a Problem in the Doctrine of Chances. By the Late Rev. Mr.* Bayes, *F.R.S. communicated by Mr.* Price, *in a letter to* John Canton, *A.M.F.R.S.*, 53 PHIL. TRANSACTIONS ROYAL SOC. LONDON 370 (1763).

[243] These estimates are also derived from an elementary statistical concept, this time, the correlation coefficient. If a = p(prejudice and IAT failure), b = p(no prejudice and IAT failure), c = p(prejudice and IAT passing) and d = p(no prejudice and IAT passing), the correlation coefficient is (ad – bc)/(a*b*c*d)^.5. If we make assumptions

true prejudice could be as low as 20% when 70% of people "fail" the IAT, the p(IAT failure| prejudice) is 90%, and the correlation of the IAT with criterion variables is .22. This latter value is virtually identical to the average correlation found in a meta-analysis of studies examining correlations between implicit and explicit attitude measures.[244] Put another way, the data invoked by IAT proponents to support the convergent validity of their test is logically consistent with the view that that the population base rate of true prejudice is far closer to survey research estimates than that implied by IAT proponents.

These statistical estimates, by themselves, neutralize much of the purported relevance of implicit prejudice research to antidiscrimination law. But we view these estimates as still too flattering to the IAT. The estimates are too generous because they rest on the shaky assumption that IAT correlations with criterion variables are perfect proxies for the IAT correlation with the hypothetical construct of true prejudice—an assumption we reject in view of the plethora of alternative explanations for such correlations noted earlier.

The crux of the problem is that correlation coefficients—which are relied on almost exclusively in the IAT literature—are not necessarily sensitive to asymmetries between test failure rates and the base rate likelihood of the phenomenon the test is supposed to assess. Even when p(prejudice | IAT failure) is low, because truly prejudiced people constitute but a small set of IAT failure cases, the resulting validity coefficient can indicate a significant relationship if we are correlating variables with markedly different base rates. As Fiedler and his colleagues note, "this problem is widely ignored in validity studies in general and in IAT research in particular."[245] This relative insensitivity of correlation coefficients also creates special difficulties for translating psychological results into legal contexts in which key concepts, such as preponderance of evidence and reasonable doubt, presuppose inferences about base rates (how pervasive are prejudicial response tendencies in this population?) and conditional probabilities (how likely is someone who acts this way to discriminate against that group, and how likely are people who discriminate against that group to behave this way?).[246]

---

about p(IAT failure), p(prejudice), and p(IAT failure | prejudice), we can solve for the correlation between IAT and prejudice. We can then restrict that range from .2 to .25, find the associated values of p(prejudice | IAT failure), and convert those values to p(no prejudice | IAT failure).

[244] *See* Hofmann et al., *supra* note 131, at 1379 ("[W]e found a small but significant positive mean population correlation of .24 between self-reported representations and representations assessed with the IAT.").

[245] Fiedler et al., *supra* note 188, at 10.

[246] The IAT data look even weaker when we add to the mix a likely bias in published IAT studies that favors publication of studies showing a significant relationship

2. *Correlation Coefficients Can Be Biased by Outliers*

The probative value of the correlations reported in IAT research is further undercut by the inferential threats posed by outliers. Implicit-prejudice researchers typically fail to check the possibility that, even if their implicit measures of prejudice do tap into a psychological construct we can fairly label prejudice and this construct does exert causal impact on criterion variables, the results may be driven by a small number of extreme scorers on either the independent or dependent variable side of the equation.[247] The

---

between IAT scores and criterion variables and disfavors publication of studies failing to reject the null hypothesis of no correlation.

Both behavioral researchers and statisticians have long suspected that the studies published in the behavioral sciences are a biased sample of the studies that are actually carried out. The extreme view of this problem, the 'file drawer problem,' is that the journals are filled with the 5% of the studies that show Type I errors, while the file drawers back at the lab are filled with the 95% of the studies that show nonsignificant (e.g., $p > .05$) results. Robert Rosenthal, *The "File Drawer Problem" and Tolerance for Null Results*, 86 PSYCHOL. BULL. 638, 638 (1979) (citations omitted). This "file-drawer problem" is not a hypothetical concern: "There is ample evidence that publication bias exists." Jack L. Vevea & Carol M. Woods, *Publication Bias in Research Synthesis: Sensitivity Analysis Using A Priori Weight Functions*, 10 PSYCHOL. METHODS 428, 428 (2005). Indeed, one of the IAT developers, Anthony Greenwald, warned long ago about the pernicious effects of publication bias. *See* Anthony G. Greenwald, *Consequences of Prejudice Against the Null Hypothesis*, 82 PSYCHOL. BULL. 1, 15 (1975) (noting that his quantitative estimates of effects of a publication bias against the null hypothesis are "frightening, even calling into question the scientific basis for much published literature"). Yet, in Greenwald and his colleagues' meta-analysis of studies examining the predictive validity of IAT scores, these authors only speculate that there is no publication bias within the IAT research, *see* Poehlman et al., *supra* note 62, at 22; they do not report any sensitivity analyses to test for the potential impact of such a bias.

[247] *See, e.g.*, John M. Orr et al., *Outlier Detection and Treament in I/O Psychology: A Survey of Researcher Beliefs and an Empirical Illustration*, 44 PERSONNEL PSYCHOL. 473, 473 (1991) ("An essential aspect of data analysis is examination of the data set to determine whether all points are appropriate for inclusion in the study at hand."). Rothermund and colleagues note that outliers may pose significant problems in the scoring of the IAT:

> Besides using a millisecond metric on the basis of correct responses, eliminating outliers is the norm rather than the exception in experimental psychology. The *D* measure [for scoring the IAT], however, is nonstandard: It is based on individually standardized reaction time differences; makes use of practice trials; does not eliminate outlier values at the right or left tail of the response time distribution; and includes reaction times based on second, correcting responses after erroneous responses or uses arbitrary error penalties. Finally, it should be emphasized that the *D* measure might be sensitive to strategic factors, because it was chosen to yield maximal correlations of IAT effects with self-report measures.

Rothermund et al., *supra* note 175, at 428 (citation omitted).

failure to conduct such sensitivity analysis would be a flaw in any research program, but the need for such analysis is particularly great for a research program that is being used to indict the vast majority of the American population as implicitly prejudiced and disposed to discriminate whenever a suitable pretext emerges.

Consider what happens when we perform rudimentary sensitivity analyses on the data collected by Ziegert and Hanges,[248] which is the only study to date to suggest that the IAT predicts discriminatory decisions in an experimental simulation of managerial decision making and which Kang and Banaji invoke as evidence for the predictive power of the IAT with respect to ultimate decisions in employment settings.[249] We find that the predictive power of the IAT falls to statistical nonsignificance when only three subjects whose scores qualify as outliers under traditional tests (scores over three standard deviations away from the mean) are deleted from the analysis (out of a total of ninety-seven subjects).[250] These three individuals were not, moreover, extreme scorers on the IAT or indeed on either of the two self-report measures of racial attitudes used in the Ziegert and Hanges study; they

---

[248] *See* Ziegert & Hanges, *supra* note 95.

[249] *See* Kang & Banaji, *supra* note 3, at 1073.

[250] *See* GREGORY MITCHELL ET AL., SENSITIVITY ANALYSIS OF ZIEGERT & HANGES 1 (2005) (manuscript on file with authors). With standardized, or Z, scores, the traditional definition of outlier is any Z score above 3 or below -3. *See* BORIS IGLEWICZ & DAVID C. HOAGLIN, HOW TO DETECT AND HANDLE OUTLIERS 10 (1993). Jonathan Ziegert and Paul Hanges kindly shared their original data with us so that we could conduct these sensitivity analyses. It should also be noted that Professors Ziegert and Hanges advocate alternative methods for outlier detection, some of which do not result in the relationship between IAT scores and ratings of hypothetical White and Black job candidates becoming non-significant. More specifically, they contend that multivariate outlier tests should be relied on rather than the simple measures discussed in the text. A number of multivariate outlier tests exists, and, under some of these tests, Ziegert and Hanges' results remain statistically significant, but not under others. However, the significant relation between IAT scores and discriminatory behavior did not hold up in robust regression analysis on the Ziegert and Hanges data (robust statistical approaches are designed to perform well even when outliers are present in the data). Furthermore, quantile regression (which allows one to make more differentiated statements about the IAT score-behavior relation within the sample) revealed that the relation between IAT scores and discriminatory behavior held only for persons showing the greatest anti-black bias on the dependent measure (i.e., for the quantile of subjects whose difference in ratings between resumes of Black and White hypothetical candidates was greatest). Our point with this illustration is not to show that the Ziegert and Hanges' study is flawed but rather to demonstrate that "more biased" IAT scores are not reliable predictors of discriminatory behavior, even in the aggregate, given the sensitivity of the relationship between IAT scores and discriminatory behavior to the influence of persons with extreme scores in the experiment.

were extreme scorers with respect to the difference between ratings for hypothetical Black and White job candidates (that is, people who assigned Blacks much lower when there was pressure from the president of the company to do so). Such results are hardly compatible with a portrait of American public opinion in which most citizens are eager to seize on the thinnest pretexts to treat Blacks unfairly. The results are far more compatible with the view that, with a few exceptions difficult to identify with any of the measures of prejudice used in the study, most people treated Blacks fairly.[251]

3. *Observed Statistical Relationships May Be Stimuli Specific*

As described previously, a subject's score on the IAT is a difference score: one scores as more prejudiced to the degree one responds faster to "compatible" pairings (e.g., White/good, Black/bad) than to "incompatible" pairings (e.g., White/bad; Black/good). This simple scoring rule has great intuitive appeal. Test-takers, especially those who fear that racism remains a powerful force in our society, can readily imagine hidden anti-Black prejudices slowing them down as they fumble about finding the left or right keys when responding to the incompatible pairings. Unfortunately, these test-takers may be beating themselves up for no good reason: a finding of "implicit bias" on the IAT may depend greatly on the particular stimuli chosen to test for this implicit bias; when the stimuli are changed, this ostensible bias may disappear.

One may have different reaction times to Black and White stimuli not because of the relative strengths of one's associations linking the very abstract categories, White and Black, to the even more abstract categories, Good and Bad, but rather because of the relative strength of: (a) mental associations or strategies that predate the experiment and that link specific White and Black stimuli in particular versions of the test to Good/Bad (e.g., the stereotypically White name "Madison" and Black name "Latonya" are linked to positive and negative economic outcomes in the real world);[252] and

---

[251] Indeed, in the Ziegert and Hanges study, after removing the three scores that we label as outliers, IAT scores did not relate to discriminatory behavior even when subjects played the role of managers specifically incited to discriminate by the hypothetical boss. *See* Ziegert & Hanges, *supra* note 95, at 559 (reporting a statistically significant relation between IAT scores and discriminatory behavior toward Blacks by subjects acting as hypothetical managers, but only in the condition in which the "boss" has specifically requested that subjects discriminate in favor of Whites in hiring).

[252] Fryer and Levitt found in their post-1970 sample of children born in California that Black parents who give their children distinctively Black names tend to be poorer than those who do not. *See* Fryer & Levitt, *supra* note 41, at 783 ("In almost all cases, variables associated with low socioeconomic status are also associated with Blacker

(b) mental associations or strategies that subjects improvise in the experimental session to help in linking the categories of White and Black, Good and Bad, or the stimuli representing those categories to the left and right response keys.[253]

This is no quibble over methodological details, for when it comes to the IAT, the devil lurks in the details. Growing evidence indicates that IAT difference scores depend greatly on the specific stimuli chosen for the sorting task. Fiedler and his colleagues cite several studies to this effect and describe a particular study in which it proved possible to invert West Germans' IAT difference scores toward East and West Germans when the specific target-related stimuli (e.g., positive West-related stimuli such as DEMOCRACY versus negative East-related stimuli such as COMMUNISM) or evaluative stimuli (e.g., traits such as SOCIABLE or IMPERSONAL) are replaced by an alternative set of "representative" stimuli.

> This raises the question of whether the IAT assesses the person's attitude toward the concept of East Germans, the meaning of the stimuli . . . , the specific Germans chosen to represent the attitude objects . . . , or the surnames and forenames as generic stimuli . . . , all of which are in principle distinct from the general concept. More generally, the crucial question is whether the IAT involves person scaling or stimulus scaling.[254]

---

names. Moreover, the link between low socioeconomic status and Black names becomes much stronger over time.").

[253] For instance, Nosek et al. found that older adults evidenced greater implicit bias toward "Black" names than photographs of Black persons. *See* Nosek et al., *supra* note 6, at 111 ("[O]lder participants tended to show stronger implicit bias against Black relative to White and old relative to young on the name versions of the task but not the faces version"); *see also* Cassandra L. Govan & Kipling D. Williams, *Changing the affective valence of the stimulus items influences the IAT by re-defining the category labels*, 40 J. EXPERIMENTAL SOC. PSYCHOL. 357, 363 (2004) ("Both studies provide support for the hypothesis that IAT effects are not solely a function of category labels. The stimulus items chosen to represent the category labels are also important, and they may drive participants to re-define the categories."). As Fiedler and his colleagues observe, current IAT scoring procedures entail "very strong assumptions" that the two classes of attitude objects (e.g., Whites and Blacks) have the same impact on final scores, that the specific stimuli chosen to represent the categories of attitude objects (selection of White and Black names) and to represent the categories of Good and Bad (selection of positive and negative words) have the same impact on final scores, and that systematic and random measurement error influences scores on the incompatible and compatible trials to the same extent. *See* Fiedler et al., *supra* note 188, at 22. Insofar as differential measurement error operates on the two components of the difference score, the zero point on the measurement scale will become increasingly meaningless.

[254] *Id.* at 14–15 (citations omitted).

The sensitivity of IAT difference scores to the specific stimuli used to represent the latent concepts of true interest (attitudes towards Whites versus Blacks or East versus West Germans) cuts to the heart of controversies discussed in our sections on construct validity:[255] does the IAT measure deep-seated prejudices toward latent concepts or does it simply rediscover the obvious semantic fact that some words, such as TORTURE and PROSPERITY, come pre-charged with negative or positive evaluative meaning? If a relative aversion toward a category only materializes when we use certain types of names or faces—or when we embed those stimuli in certain types of context cues—are we justified in saying people are prejudiced toward the category as a whole or only toward particular exemplars of that category in certain contexts?

And it is important to note that most of the IAT data derives from a fairly limited set of stimuli given the redundant use of the most common IATs. IAT researchers boast that they have collected the results of millions of IATs.[256] However, while the sample of subjects is large and diverse, the sample of stimuli is not nearly as impressive. Not too long ago one of social cognition's most popular biases, the overconfidence effect, was found to be in many cases the by-product of the sample of experimental stimuli repeatedly used to demonstrate the bias.[257] We suspect the same fate may well befall many of the implicit "biases" demonstrated with the various IATs.

D. *External Validity and the Law's Requirement of "Fit": Unpacking the* Ceteris Paribus *Clause Attached to the IAT Effect*

External validity refers to the degree to which one can generalize causal or statistical relationships found by relying on particular measures of particular subjects in a particular setting to other possible measures, subjects and settings. This last set of validity challenges is perhaps the most devastating for the application of implicit prejudice research to the law,[258]

---

[255] *See supra* Parts III.A.3–4.

[256] *See, e.g.*, Kang & Banaji, *supra* note 3, at 1072 (referencing a database containing "well over three million tests").

[257] *See* Peter Juslin, *Representative Design: Cognitive Science from a Brunswikian Perspective*, *in* THE ESSENTIAL BRUNSWIK: BEGINNINGS, EXPLICATIONS, APPLICATIONS 404, 404–08 (Kenneth R. Hammond & Thomas R. Stewart eds., 2001); Peter Juslin, Anders Winman & Henrik Olsson, *Naive Empiricism and Dogmatism in Confidence Research: A Critical Examination of the Hard-Easy Effect*, 107 PSYCHOL. REV. 384, 384–85, 387–88 (2000).

[258] *Cf.* Vladimir J. Konečni & Ebbe B. Ebbesen, *External Validity of Research in Legal Psychology*, 3 LAW & HUMAN BEHAV. 39, 40 (1979) ("Research in applied disciplines must be concerned with issues of external validity and generalizability to an

for, even if the implicit prejudice research program had emerged unscathed from the previous three sets of tests of validity, the obstacles to generalizing from the typical implicit prejudice research paradigm to realistic workplace settings are daunting.

Legal scholars promoting implicit prejudice research employ far too simplistic a general-causation model linking stereotypes and racially-biased personnel judgments. They rarely make more than superficial and selective reference to the long list of personality and situational factors that can influence, sometimes dramatically, the degree to which, and even the direction in which, stereotypes can bias judgments.[259] Moreover, the omissions conveniently favor assumptions of implicit prejudice as an implacable causal force. Once we start acknowledging these omissions, we must also start acknowledging the psychological importance of the many ways in which workplaces and other discrimination settings systematically differ from the artificial settings so common in implicit prejudice research.

Consider the following dozen specific ways in which work settings (many modeled on best-practices precepts in organizational behavior)[260]

---

unusually high degree. The criteria for what is a good experiment, when a certain methodology appears sound, and which results are to be trusted, must necessarily be different and more stringent when sweeping, costly, and far-reaching changes in public policy and—quite literally in the legal domain—people's futures and lives, depend on inferences from research results.").

[259] *See, e.g.*, Kang, *supra* note 16, at 1567 ("The social cognition studies that I have presented are not without ambiguity, confusion, and contradiction. They often raise as many questions as they answer. That said, a prima facie case has been made about the existence of implicit bias, its dissociation from explicit self-reports of bias, its measurability through reaction time designs, and its impact on behavior."). Kang and Banaji dismiss external validity concerns simply by citing one unpublished *experiment* that used medical interns as subjects (rather than the usual convenience sample of college undergraduates) and found some possible connection between the interns' IAT scores and their differential recommendations for treatment of hypothetical White versus Black patients. *See* Kang & Banaji, *supra* note 3, at 1072. A few of the advocates of implicit bias research within the behavioral realist movement do forthrightly acknowledge the limited predictive ability of this research. *See* Krieger & Fiske, *supra* note 25, at 91 ("In short, whether attitudes predict behavior depends on the attitude, the context, and the person .... Specifically, bias, understood as an attitude toward members of a particular group, predicts overt discrimination only under some circumstances."); *id.* at 103 ("[C]ontext is a powerful moderator of implicit bias.").

[260] For an itemization of oft-cited best practices in checking prejudice, see generally MICHAEL ARMSTRONG, A HANDBOOK OF HUMAN RESOURCE MANAGEMENT PRACTICE (9th ed. 2003); *see also* U.S. Equal Employment Opportunity Commission, Best Practices of Private Sector Employees, http://www.eeoc.gov/abouteeoc/task_reports/practice.html (last visited Jan. 11, 2006). Although many major companies implement several of these practices, our knowledge remains thin on: (a) how widespread and carefully implemented these checks are; and (b) how many checks on superficial biased processing is enough.

differ from the typical laboratory experiment on stereotyping and prejudice:

(1) Subjects in laboratory experiments are typically asked to judge people about whom they have virtually no work-history or past-performance information (indeed, subjects often have little more information than group category membership, like race or gender, on which to base their impressions); however, hiring and staffing managers often have access to a great deal of carefully compiled data on the past behavior and performance of those they are judging;

(2) Subjects in laboratory experiments are not typically judging people with whom they expect to interact in the future; yet, hiring and staffing managers often expect future interaction (including potentially awkward social encounters in which they must explain in detail to those whom they have judged why the ratings or outcomes take the form they do);

(3) Subjects in laboratory experiments are not typically asked to judge people whom they perceive to be on their "team" (people with whom they must work collaboratively to achieve shared goals—an independent variable some psychologists have designated as "outcome interdependence"); hiring and staffing managers have strong interests in choosing the best possible people because the people they choose will indeed be on their team and potentially affect their own future performance evaluations and raises;

(4) Subjects in laboratory experiments rarely, if ever, have powerful long-term financial or legal incentives for doing a better job at performance appraisal; hiring and staffing managers typically have strong financial, legal, and long-term incentives for making correct and lawful decisions;

(5) Subjects in laboratory experiments are rarely given any training in using rating scales or other evaluation tools; hiring and staffing managers at many organizations receive extensive training in performance evaluation and making compensation decisions, and these persons have been alerted to the dangers of a variety of rating biases as well as the dangers of discrimination;

(6) Subjects in laboratory experiments rarely expect that they will have to justify their opinions to people above them in an organizational hierarchy who control important rewards and punishments; hiring and staffing managers are well aware of the need to have adequate and legal justifications for their judgments and decisions;

(7) Subjects in laboratory experiments rarely expect that they will be accountable to high-status others who have repeatedly affirmed a non-

discrimination policy; hiring and staffing managers at many organizations are often aware of the views of those to whom they are accountable and of the high value that is placed on achieving diversity goals and avoiding charges of discrimination;

(8) Subjects in laboratory experiments rarely receive clear or repeated admonishment not to allow job-performance-irrelevant characteristics, such as membership in ethnic-racial groups, to affect their personnel judgments; hiring and staffing managers often receive clear and repeated admonishment on this score;

(9) Subjects in laboratory experiments are rarely encouraged or required to attend training workshops on how to make personnel decisions; the opposite is true of hiring and staffing managers at large organizations, as well as many small to mid-size organizations;

(10) Subjects in laboratory experiments are rarely given written manuals and guidelines that place constraints on how they should perform their task; the opposite is true of hiring and staffing managers at many large organizations, as well as many small to mid-size organizations;

(11) Subjects in laboratory experiments rarely expect that they will have to offer comparative rationales for why they selected certain people and did not select others; hiring and staffing managers are often expected to do so; and

(12) Subjects in laboratory experiments are typically college undergraduates who have had virtually no experience supervising and evaluating employees; hiring and staffing managers typically are fully mature adults who have considerable experience in supervisory roles.

Entire chapters of the most recent edition of the *Handbook of Social Psychology* have been devoted to personality, cultural, and organizational threats to the generalizability of experimental findings,[261] and these external

---

[261] *See* Robert B. Cialdini & Melanie R. Trost, *Social Influence: Social Norms, Conformity, and Compliance*, *in* 2 HANDBOOK OF SOCIAL PSYCHOLOGY, *supra* note 33, at 151; Alan Page Fiske et al., *The Cultural Matrix of Social Psychology*, 2 HANDBOOK OF SOCIAL PSYCHOLOGY, *supra* note 33, at 915; John M. Levine & Richard L. Moreland, *Small Groups*, *in* 2 HANDBOOK OF SOCIAL PSYCHOLOGY, *supra* note 33, at 415; Jeffrey Pfeffer, *Understanding Organizations: Concepts and Controversies*, *in* 2 HANDBOOK OF SOCIAL PSYCHOLOGY, *supra* note 33, at 733; Mark Snyder & Nancy Cantor,

validity constraints counsel against sweeping causal claims such as the contention that subjectivity in personnel decision-making opens the door wide to stereotyping and prejudice.[262] An inspection of the business school curriculum, and of employment practices in Fortune 500 companies, would reveal the implementation of many of the strategies these *Handbook* chapters recommend to check cognitive and motivational biases.[263] Worse still for those hoping that these objections are merely hypothetical (that is, of the form, "these differences could make a difference—although there is no evidence that they do"), it is possible to identify experimental studies that manipulate many of the contextual factors contained in the list of the dozen differences. In other words, experimental research has repeatedly shown that these differences do make a difference.

For instance, it has long been known that the likelihood of evaluators falling prey to a wide range of cognitive biases can be affected by such factors as the degree to which: (a) these evaluators expect to be called on to justify their performance assessments and decisions to third parties;[264] (b) evaluators have job-relevant information—a factor on which there is professional consensus among industrial-organization psychologists;[265] (c)

Understanding Personality and Social Behavior: A Functionalist Strategy, *in* 1 HANDBOOK OF SOCIAL PSYCHOLOGY, *supra* note 33, at 635.

[262] *See* Bielby, *supra* note 109, at 381 ("In decision-making contexts characterized by arbitrary and subjective criteria and substantial decision-maker discretion, individuals tend to seek out and retain stereotype-confirming information and to ignore or minimize information that defies stereotypes." (footnote omitted)); *see also* Hart, *supra* note 151, at 745 ("There is little doubt that unconscious discrimination plays a significant role in decisions about hiring, promoting, firing, and other benefits and tribulations in the workplace.").

[263] *See* Arthur & Doverspike, *supra* note 32, at 307–14; *see also* Cora Daniels, *50 Best Companies for Minorities*, FORTUNE, June 28, 2004, at 136–38.

[264] *See* Jennifer S. Lerner & Philip E. Tetlock, *Accounting for the Effects of Accountability*, 125 PSYCHOL. BULL. 255, 258 (1999). And the desire to get along with the other person should be an especially effective control on expressions of prejudice when the other person is a superior believed to be hostile to discrimination. *See* Stacey Sinclair et al., *Social Tuning of the Self: Consequences for the Self-Evaluations of Stereotype Targets*, 89 J. PERSONALITY & SOC. PSYCHOL. 160, 172 (2005) (indicating that "these experiments showed that individuals' self-evaluations and behavior shifted in response to the ostensible views of short-term interaction partners as a function of affiliative motivation").

[265] *See* Michele J. Gelfand et al., *Discrimination in Organizations: An Organizational-Level Systems Perspective*, *in* DISCRIMINATION AT WORK: THE PSYCHOLOGICAL AND ORGANIZATIONAL BASES, *supra* note 32, at 89, 101 ("The best way to combat discrimination in selection is to use measures that tap as many aspects of job performance as possible, to utilize different media in terms of the ways in which content is presented and responses are required . . . , and to use noncognitive measures such as personality and integrity tests when possible." (citation omitted)).

evaluators interact with individuals whose behavior runs counter to, or is not readily assimilable into, the stereotype;[266] (d) evaluators work in settings in which race is not a useful predictive cue;[267] (e) evaluators have ready access to "individuating" information that can "dilute" the impact of group-based stereotypes;[268] (f) evaluators expect future interactions with those they are judging;[269] (g) evaluators feel functionally interdependent with those they are judging (e.g., on the same team);[270] (h) evaluators feel that they have financial or other incentives for getting the answer right;[271] (i) evaluators are highly motivated to control prejudicial reactions and recognize the risk of prejudice in the immediate situation;[272] and (j) evaluators work in

---

[266] *See* Christopher L. Aberson et al., *Implicit Bias and Contact: The Role of Interethnic Friendships*, 144 J. SOC. PSYCHOL. 335 (2004); Irene V. Blair, *The Malleability of Automatic Stereotypes and Prejudice*, 6 PERSONALITY & SOC. PSYCHOL. REV. 242 (2002).

[267] *See* Plant et al., *supra* note 146, at 153–54.

[268] *See* RICHARD NISBETT & LEE ROSS, HUMAN INFERENCE: STRATEGIES AND SHORTCOMINGS OF SOCIAL JUDGMENT 154–56, 267–68 (1980); Philip E. Tetlock & Richard Boettger, *Accountability: A Social Magnifier of the Dilution Effect*, 57 J. PERSONALITY & SOC. PSYCHOL. 388 (1989); *see also* Copus, *supra* note 147, at 453 ("In those real-world employment decisions the parties involved have had extensive personal contact over an extended period, a critical condition that few if any laboratory studies have even attempted to replicate. The decision makers have had extended opportunity to observe the employees and gather much information relevant to the decision at hand—diagnostic individuating information, in the jargon of psychologists." (citation omitted)).

[269] *See* Barry R. Schlenker et al., *Coping with Accountability: Self-Identification and Evaluative Reckonings*, *in* HANDBOOK OF SOCIAL & CLINICAL PSYCHOLOGY: THE HEALTH PERSPECTIVE 96 (C.R. Snyder & Donelson R. Forsyth eds., 1991); *see also* Copus, *supra* note 147, at 453 ("[T]he decision maker typically anticipates having further interaction with the target employees.").

[270] *See* Susan T. Fiske, *Thinking Is for Doing: Portraits of Social Cognition from Daguerreotype to Laserphoto*, 63 J. PERSONALITY & SOC. PSYCHOL. 877, 879 (1992); Susan T. Fiske & Steven L. Neuberg, *A Continuum of Impression Formation, from Category-Based to Individuating Processes: Influences of Information and Motivation on Attention and Interpretation*, 23 ADVANCES IN EXPERIMENTAL SOC. PSYCHOL. 1 (1990).

[271] *See* Ralph Hertwig & Andreas Ortmann, *Experimental Practices in Economics: A Methodological Challenge for Psychologists?*, 24 BEHAV. & BRAIN SCI. 383, 391 (2001); Erik P. Thompson et al., *Accuracy Motivation Attenuates Covert Priming: The Systematic Reprocessing of Social Information*, 66 J. PERSONALITY & SOC. PSYCHOL. 474 (1994).

[272] *See* Bridget C. Dunton & Russell H. Fazio, *An Individual Difference Measure of Motivation to Control Prejudiced Reactions*, 23 PERSONALITY & SOC. PSYCHOL. BULL. 316, 324–25 (1997); Duane T. Wegener & Richard E. Petty, *The Flexible Correction Model: The Role of Naïve Theories of Bias in Bias Correction*, 29 ADVANCES IN EXPERIMENTAL SOC. PSYCHOL. 141 (1997); Wheeler & Petty, *supra* note 197; *see also*

cooperative and egalitarian settings that "cue" appropriate behavior.[273]

These lines of work, almost always absent from recent law review scholarship on unconscious discrimination, warn us that the differences between laboratory settings and applied domains significantly constrain the generalizability of implicit prejudice research. The results repeatedly show that even pallid laboratory imitations of the many checks on prejudice can attenuate biased thinking, inducing people to question their preconceptions and to become more attuned to the limitations of their own knowledge.

Furthermore, our list of twelve external validity concerns far from exhausts the external validity challenges. Five other lines of social-scientific research suggest that several scholars have exaggerated the threat of automatically activated unconscious stereotypes and prejudice. These lines of work demonstrate:

(a) the automatic activation of stereotypes of a racial group by the visual presence of a member of that group is not as automatic as often implied; context matters.[274] The preponderance of evidence now indicates that photos of Black persons embedded in an egalitarian and positive-affect setting do not evoke the millisecond response-time differentials that some social psychologists take as evidence of prejudice and, for a non-negligible number of Whites, such positively embedded photos evoke "pro-Black" responses.[275]

(b) both contrast and assimilation effects occur. This means that if a White and a Black person each behave exactly the same way (e.g., both are hard working) and that behavior contradicts a negative stereotype that observers hold of Black people, observers sometimes react *more, not less*, favorably to the Black person.[276]

Copus, *supra* note 147, at 453 ("[M]ost companies make extensive efforts to make equal opportunity and nondiscrimination a salient goal in all employment decisions.").

[273] *See* Patricia G. Devine et al., *The Regulation of Explicit and Implicit Race Bias: The Role of Motivations to Respond Without Prejudice*, 82 J. PERSONALITY & SOC. PSYCHOL. 835 (2002); Margo J. Monteith et al., *Putting the Brakes on Prejudice: On the Development and Operation of Cues for Control*, 83 J. PERSONALITY & SOC. PSYCHOL. 1029, 1046–47 (2002); Rudman et al., *supra* note 90.

[274] *See* Lowery et al., *supra* note 131, at 851 ("Taken together, the four experiments presented here provide evidence that automatic racial attitudes are subject to both tacit and explicit social influence . . . . [O]ur results contradict the assumption that automatic prejudice is inevitably activated by the presence of a member of a negatively stereotyped group, and thereby may represent a challenge to existing theories of automatic stereotyping and prejudice.").

[275] *See* Jamie Barden et al., *Contextual Moderation of Racial Bias: The Impact of Social Roles on Controlled and Automatically Activated Attitudes*, 87 J. PERSONALITY & SOC. PSYCHOL. 5, 16, 21 (2004).

[276] *See* ZIVA KUNDA, SOCIAL COGNITION: MAKING SENSE OF PEOPLE 351–69 (1999); Ziva Kunda & Steven J. Spencer, *When Do Stereotypes Come to Mind and When Do They Color Judgment? A Goal-Based Theoretical Framework for Stereotype Activation*

(c) stereotype effects recede as people learn more about each other as individuals, with individuating information often overwhelming stereotype information. Kunda and Spencer note that even if stereotypes are activated at the start of an interaction, this activation can dissipate quickly: "In more than half a dozen studies, we have found no trace of stereotype activation in participants who had observed or interacted with a Black or an Asian individual for about [10 minutes]."[277]

(d) intergroup contact is associated with reduced prejudice,[278] including "prejudice" at the implicit level.[279] Further, "contact theory applies beyond racial and ethnic groups to embrace other types of groups as well."[280]

(e) the strong form of the "subjectivity-opens-the-door-to-prejudice" hypothesis is strikingly disconfirmed by recent meta-analyses of diverse datasets bearing on ethnic group differences in job performance. These meta-analyses, published in the most selective journal in industrial-organizational psychology, show that objective measures of job performance are associated with very similar, and sometimes larger, standardized ethnic group differences than are subjective measures.[281]

---

and Application, 129 PSYCHOL. BULL. 522 (2003); Leonard L. Martin et al., *Assimilation and Contrast as a Function of People's Willingness and Ability to Expend Effort in Forming an Impression*, 59 J. PERSONALITY & SOC. PSYCHOL. 27 (1990).

[277] Kunda & Spencer, *supra* note 276, at 523; *see also* Lee Jussim et al., *Why Study Stereotype Accuracy and Inaccuracy?*, *in* STEREOTYPE ACCURACY: TOWARDS APPRECIATING GROUP DIFFERENCES 3, 13 (Y.T. Lee et al., eds., 1995) ("When individuating information is ambiguous or difficult to detect, people often rely on their stereotypes rather than individuating information. However, of all the studies that have manipulated both group information . . . and the personal characteristics of targets . . . , we are not aware of a single one that has shown that people *ignore* individual differences. Perceivers base their judgments far more on the personal characteristics of targets than on targets' gender or membership in ethnic groups . . . ." (citations omitted)).

[278] *See* Thomas F. Pettigrew & Linda R. Tropp, *A Meta-Analytic Test of Intergroup Contact Theory*, 90 J. PERSONALITY & SOC. PSYCHOL. 751, 766 (2006) ("The meta-analytic results clearly indicate that intergroup contact typically reduces intergroup prejudice.").

[279] *See* Andreas Olsson et al., *The Role of Social Groups in the Persistence of Learned Fear*, 309 SCI. 785, 786 (2005) (finding reduced implicit racial bias in persons with more interracial dating experience).

[280] Pettigrew & Tropp, *supra* note 278, at 768.

[281] *See* Patrick F. McKay & Michael A. McDaniel, *A Reexamination of Black-White Mean Differences in Work Performance: More Date, More Moderators*, 91 J. APPLIED PSYCHOL. 538, 548 (2006) ("Evidence provided in Table 5 suggests that measurement method has a relatively low impact on mean racial differences in work performance (R = .10). Summary results for this moderator presented in Table 2 support this conclusion because effect sizes are very similar for subjective (d = 0.28) and objective (d = 0.22) measures of performance."); Roth et al., *supra* note 151, at 702 ("Our results do not

All of the factors listed above undercut the sweeping general-causation arguments recently featured in prominent law reviews. Failure to appreciate that these factors often make a difference leads to unwarranted optimism about the external validity of implicit prejudice research—or rather, unwarranted pessimism about the discriminatory effects of implicit prejudice as identified in experimental studies. If the task of inferring generally applicable causal laws from the implicit prejudice experiments is daunting, the task of inferring specific-causation conclusions in a specific case is no easier.[282] Imagine that an employer utilizes the IAT as a screening mechanism for the human resources staff, as some legal scholars have suggested.[283] Based on what we know about the construct and external validity constraints on IAT results, drawing conclusions about an employee's prejudicial tendencies in the workplace from IAT scores would border on recklessness. We just do not yet know enough about the sources of the IAT effect, the behavioral concomitants of this effect, and the parameters on this effect to make intelligent assessments of the workplace relevance of implicit prejudice.[284]

---

support the position that subjective measures have more potential for bias than objective measures. Instead, we found the opposite.").

[282] Further support for our skeptical stance comes from one of the leading theorists in the field of implicit social cognition, which encompasses the IAT work. John Bargh is not convinced that the field has adequate theoretical answers to questions of the general form: How do we know which of the multitude of possible behavioral scripts will be activated by a complex naturalistic "stimulus" (such as a Black or female co-worker)? How do we know which scripts will predominate? And how do we reconcile the automaton-like view of human nature that emerges from stylized priming studies with what we know about the resourcefulness with which human beings pursue complex sets of goals in natural environments fraught with obstacles (such as the world of work)? *See generally* John A. Bargh, *What Have We Been Priming All These Years? On the Development, Mechanisms, and Ecology of Nonconscious Social Behavior*, 36 EUR. J. SOC. PSYCHOL. 147 (2006). Bargh's abstract critique dovetails well with our much more focused one: the implicit prejudice research program is a very incomplete project.

[283] *See supra* notes 103–04 and accompanying text.

[284] Our point in this section is not that real organizations achieve perfection at preventing or eliminating discrimination. Field work by Pager and Quillian, for instance, suggests that racial discrimination in entry-level hiring by low-wage employers may occur with some frequency. *See* Devah Pager & Lincoln Quillian, *Walking the Talk? What Employers Say Versus What They Do*, 70 AM. SOC. REV. 355, 366 (2005) ("[A]ctual behavioral measures in the audit show that white ex-offenders are more than three times as likely to receive consideration from employers as black ex-offenders." (footnote omitted)). Rather, our point is that measures of implicit prejudice tested in lab settings have not been shown to predict discrimination in any settings approximating real world workplaces, much less workplaces that institute anti-discrimination safeguards or at which interracial interactions are common.

IV. CONCLUSION

If we accepted at face value the most ambitious claims about the pervasiveness and potency of unconscious prejudice, then the factual case for more aggressive government intervention to fight discrimination in a wide range of domains would be strengthened.[285] But scholars advancing these claims, in their eagerness to incorporate cutting-edge empirical research into their policy analyses, fail to acknowledge the complexity of the questions still surrounding the implicit prejudice research program. To recapitulate, much murkiness surrounds (a) the proper causal explanation for alleged IAT effects, (b) the psychological meaning of IAT scores, (c) the statistical generality and potency of alleged relations between IAT scores and actual behavior, and (d) boundary conditions on alleged IAT effects.

Many judges and legal scholars have learned to be wary of social scientists bearing intellectual gifts,[286] but there are special properties of the debate on implicit prejudice research to which it is prudent to call attention. Work on implicit prejudice is not just psychological; it is suffused with political significance. The concept of implicit prejudice straddles the is-ought boundary that has traditionally separated facts from values: descriptive scientific claims about how people think from normative moral-political claims about how people should think. This blurring manifests itself in many ways, but most importantly, it manifests itself in the repeated failure of

---

[285] But one would not be logically obliged to draw any conclusions about how the law should change even if one fully accepted the most far-reaching empirical claims. Much would still hinge on empirical estimates of the consequences, intended and unintended, of the new legal regime and even more still would hinge on one's moral-political values and one's willingness to make egalitarianism a trump value. *See* Krieger & Fiske, *supra* note 25, at 18 ("[E]ven the best insights from the empirical social sciences can not supply the normative principles needed for substantive lawmaking or resolve the conflicts between competing norms and interests so often implicated in the legislative and judicial processes.").

[286] *See, e.g.*, RICHARD A. POSNER, OVERCOMING LAW 526 (1995) ("Scientists seek to bolster their authority by affectations of mathematical rigor, by use of an intimidating jargon, by suppressing doubts, and by concealing the personal, judgmental factor in the evaluation of experimental, statistical, or observational results."); David L. Bazelon, *Risk and Responsibility*, 205 SCI. 277, 277 (1979) ("We are no longer content to delegate the assessment of and response to risk to so-called disinterested scientists. Indeed, the very concept of objectivity embodied in the word disinterested is now discredited."); *id.* at 278 ("[S]cientists are tempted to disguise controversial value decisions in the cloak of scientific objectivity, obscuring those decisions from political accountability."); Mark S. Brodin, *Behavioral Science Evidence in the Age of* Daubert*: Reflections of a Skeptic*, 73 U. CIN. L. REV. 867, 870 (2005) ("[A]lthough social scientists have made significant contributions to society, as suppliers of courtroom evidence they often stand on shaky ground.").

researchers to acknowledge the role that political values unavoidably play in where they set their thresholds of proof for calling reaction times or puzzling behavior evidence of prejudice. The resulting distortions help to explain widespread interpretive over-reaching: the willingness to claim revolutionary discoveries well before ruling out alternative, more pedestrian accounts of what implicit measures of prejudice assess.

The claims of implicit prejudice researchers would be less contentious if they had shown that implicit measures of prejudice correlate highly with behavior that observers across the political spectrum agree represent intergroup hostility or discrimination.[287] But implicit prejudice researchers claim construct validity for the IAT—as an implicit measure of *prejudice*[288]—by showing that IAT scores correlate weakly to moderately with eye blinks, millisecond differentials in reaction times, and other subtle behaviors that are far from universally regarded as signs of prejudice. Thus, we can accept the claims of implicit prejudice researchers that they are studying *prejudice* and not other psychological constructs only if we are willing to redefine what prejudice requires in terms of real-world referents.

This debate should not be dismissed as an esoteric feud among psychological insiders about the proper technical definitions of prejudice; it is a debate about whether social psychologists are entitled to co-opt a value-laden concept to advance their policy agenda. *Prejudice* and *racism* are not the sorts of value-neutral descriptive terms one would expect to encounter in the data language of a positivist science committed to the dispassionate weighting of rival hypotheses. These terms are not on an epistemological par with other terms that could be used to describe the concepts at issue, such as *spreading activation networks*,[289] *amygdala activation*,[290] or *response time latency*.[291] Unlike a claim that pictures of Black and White faces

---

[287] But they have not, as discussed below. *See supra* Part III.A.2. At best, measures of implicit prejudice have been shown to relate to some micro-level behaviors that can, in a technical sense, be labeled "discriminatory behaviors." Little evidence presently exists tying implicit prejudice measures to macro-level discriminatory behaviors.

[288] Or, as Banaji and colleagues put it, "[g]enuine, 100% [p]rejudice." Banaji et al., *supra* note 29, at 280.

[289] Which refers to activation of one node within a network leading to the activation of associated nodes within the network. *See, e.g.*, Charles M. Judd et al., *Some Dynamic Properties of Attitude Structures: Context-Induced Response Facilitation and Polarization*, 60 J. PERSONALITY & SOC. PSYCHOL. 193, 197–200 (1991) (presenting a spreading-activation model of attitude accessibility).

[290] Which refers to activation of "a subcortical structure known to be involved in emotional learning, memory, and evaluation." Phelps et al., *supra* note 193, at 729.

[291] Which refers to the amount of time it takes to react to a stimulus, as in the IAT. *See supra* Part II.B.

differentially activate the amygdala region within the brain,[292] charges of prejudice and racism carry powerful connotative as well as denotative judgments: such charges imply that the investigators are condemning as well as describing the attitude in question. Prejudice and racism are political hot potatoes that partisans are quick to disavow for themselves and attribute to their adversaries.[293] Simply questioning the rigor of prejudice research can subject one to questions about one's awareness of current social problems and one's commitment to equality.[294] If social psychologists insist on importing value-laden constructs from political debate into their scientific work, they should expect political fissures to ripple through the scientific debate. They should expect those convinced that social inequality flows from anti-Black prejudice in the here and now to employ wide-net measures of racism, such as opposition to affirmative action or greater eye blinking in the presence of Black versus White experimenters, that label as many people as possible prejudiced. And they should expect those convinced that inequality is now largely the result of processes internal to Black communities to view racism as more excuse than explanation and dismiss all but the starkest evidence of disparate treatment driven by racial animus.

Our fear is that the stage has been set for an epistemic disaster of minor-epic proportions. Throughout this Article, we have seen how rarely IAT researchers temper their enthusiasm for ferreting out unconscious prejudice with offsetting concerns about the dangers of making false accusations of prejudice.[295] It is beyond the scope of this Article—and indeed of science—

---

[292] William A. Cunningham et al., *Separable Neural Components in the Processing of Black and White Faces*, 15 PSYCHOL. SCI. 806, 811 (2004) ("[W]e found greater amygdala activation for Black than White faces when faces were presented for only 30 ms.").

[293] *See* Tetlock, *supra* note 60, at 526–27.

[294] In a prior exchange, Banaji et al. liken Arkes and Tetlock's criticisms of implicit prejudice research to criticisms that members of the *Plessy v. Ferguson* Court would have made of the *Brown v. Board of Education* decision. *See* Banaji et al., *supra* note 29, at 281. And Sears likens the criticisms by Arkes and Tetlock to the views of William Parker, a Los Angeles police chief in the 1960s often accused of racism. *See* Sears, *supra* note 182, at 294. More recently, in an exchange with Blanton and Jaccard, Greenwald and colleagues speculated on the defensive motives of the IAT's critics, *see* Anthony G. Greenwald et al., *Consequential Validity of the Implicit Association Test: Comment on Blanton and Jaccard* (2006), 61 AM. PSYCHOL. 56, 60 (2006)—remarks much in the spirit of Kang and Banaji's portrayal of skeptics as self-deceiving hypocrites, *see* Kang & Banaji, *supra* note 3, at 1065. For our part, we believe that what matters is the quality of the science, not the alleged motives of individual scientists.

[295] Such dogged commitment to a theory or research program may admittedly advance basic science in some cases, *see* Philip Kitcher, *The Division of Cognitive Labor*, 87 J. PHIL. 5, 8 (1990) ("Whereas it may be rational for each of the scientists to believe

to stipulate how society should balance Type I and Type II errors in the courtroom and political arena. It must suffice to note that both errors are serious—so serious that a balanced social science cannot focus exclusively on only one of these errors.[296] If the knowledge claims of IAT advocates are as exaggerated as we maintain, IAT advocates are already causing substantial harm to American society by: (a) stimulating excessive suspicion of Whites among Blacks, suspicion that can crystallize into conspiracy theories that poison race relations;[297] (b) convincing Blacks that they are held in contempt, thereby inducing "stereotype threat" and "social-identity threat" that, respectively, increase the likelihood of self-fulfilling prophecies in which Blacks act in ways that confirm the ill opinions they imagine others hold[298] and heighten preconscious attention to subtle cues that confirm the devalued role of minority groups;[299] (c) providing authoritative-sounding but false feedback to a million-plus visitors to IAT websites that they are

---

the theory that is better supported by the available evidence, it may not be rational for each of them to pursue that theory, and what the community cares about is the distribution of pursuit not the distribution of belief."), but drawing applications from such developing research programs may be disastrous.

[296] *See* Robert J. MacCoun, *Biases in the Interpretation and Use of Research Results*, 49 ANN. REV. PSYCHOL. 259, 273–75 (1998) (discussing the importance of considering standards for avoiding false negatives versus false positives in the evaluation of a scientific program of research); *see also* Daubert v. Merrell Dow Pharm., Inc., 509 U.S. 579, 594 (1993) ("[I]n the case of a particular scientific technique, the court ordinarily should consider the known or potential rate of error and the existence and maintenance of standards controlling the technique's operation." (citations omitted)).

[297] The potential harm in exaggerating beliefs about the pervasiveness of implicit prejudice within the American public is illustrated by research demonstrating that "the more ethnic minorities expected whites to be prejudiced, the more negative experiences they had during interethnic interactions." *See* J. Nicole Shelton et al., *Expecting To Be the Target of Prejudice: Implications for Interethnic Interactions*, 31 PERSONALITY & SOC. PSYCHOL. BULL. 1189, 1189 (2005); *see also* PAUL M. SNIDERMAN & THOMAS PIAZZA, BLACK PRIDE AND BLACK PREJUDICE 50–53 (2002) (reporting on a national survey indicating that the more Blacks believed prejudice among Whites to be pervasive, the more likely they were to embrace conspiracy theories for minority inequality).

[298] *See, e.g.*, Claude M. Steele & Joshua Aronson, *Stereotype Threat and the Intellectual Test Performance of African Americans*, 69 J. PERSONALITY & SOC. PSYCHOL. 797, 808 (1995) ("[M]aking African American participants vulnerable to judgment by negative stereotypes about their group's intellectual ability depressed their standardized test performance relative to White participants, while conditions designed to alleviate this threat, improved their performance."). For a critical discussion of the stereotype threat phenomenon, see generally Paul R. Sackett et al., *On Interpreting Stereotype Threat as Accounting for African American-White Differences on Cognitive Tests*, 59 AM. PSYCHOL. 7 (2004).

[299] Cheryl R. Kaiser et al., *Prejudice Expectations Moderate Preconscious Attention to Cues that Are Threatening to Social Identity*, 17 PSYCHOL. SCI. 332, 337 (2006).

prejudiced;[300] and (d) providing authoritative-sounding but false grounds for commonality-of-cause requirements in class action litigation.[301]

Empirical claims that carry serious policy implications require serious scrutiny—and the more sweeping the claims, the heavier the burden of proof their promoters should bear. And it is a sweeping claim to say that, after half a century of legal, political and educational efforts to check prejudice, the vast majority of Americans remain prejudiced. When psychological and legal scholars join forces to call for wholesale changes in American antidiscrimination law on the basis of this implicit-prejudice charge, more is at stake than professorial reputations. These claims from the left shake the ontological foundations of American political culture as profoundly as do claims from the right about the genetic causes of intelligence and income inequality.[302] The former claims imply that we cannot achieve equality of opportunity unless we have already achieved equality of result (otherwise people will implicitly associate subordinate group membership with bad outcomes in the lottery of life—and those associations will bias gatekeepers against members of those groups); the latter claims imply that we cannot achieve equality of opportunity without achieving an even more fundamental equality: equality of DNA that shapes our minds and personalities. Either way, the American dream is vastly more elusive than popularly supposed.

Without denying that disturbing claims, from left or right, sometimes prove on the mark, we believe that psychologists have an obligation to subject the claims of implicit prejudice researchers to as vigorous scientific accountability as they do research on the heritability of intelligence and other

---

[300] *See* Blanton & Jaccard, *supra* note 130, at 69. Shelby Steele's work suggests that information such as that presented on the IAT website may fuel "white guilt," which makes it hard to have open conversations about current impediments to improving living standards for African-Americans. *See* SHELBY STEELE, WHITE GUILT: HOW BLACKS AND WHITES TOGETHER DESTROYED THE PROMISE OF THE CIVIL RIGHTS ERA 34–36 (2006).

[301] *See supra* notes 110–11 and accompanying text. As David Copus points out, social psychological experts are often used by plaintiffs in employment class actions to testify that psychological findings, paired with a particular organizational structure, will create an environment conducive to discrimination. E-mail from David Copus to Gregory Mitchell & Philip E. Tetlock (Dec. 26, 2005, 02:57 EST) (on file with authors).

[302] *See generally* RICHARD J. HERRNSTEIN & CHARLES MURRAY, THE BELL CURVE: INTELLIGENCE AND CLASS STRUCTURE IN AMERICAN LIFE (1994). For a discussion of psychologists' responses to *The Bell Curve*, see Clayton P. Alderfer, *The Science and Nonscience of Psychologists' Responses to* The Bell Curve, 34 PROF. PSYCHOL.: RES. & PRAC. 287 (2003). For a discussion of responses by philosophers of science, see Neven Sesardic, *Philosophy of Science that Ignores Science: Race, IQ and Heritability*, 67 PHIL. SCI. 580 (2000).

attributes conducive to success in competitive market economies.[303] Anything less would suggest that this branch of psychology is better classified as a form of social activism than of science.

---

[303] The correct scientific model is adversarial collaboration, in which the scientists, rather than impugning each other's motives and refuting caricatures of each other's positions, identify the points on which they agree and disagree and the findings that would induce them to change their minds. *See, e.g.*, Daniel Kahneman, *Experiences of Collaborative Research*, 58 AM. PSYCHOL. 723, 729–30 (2003). We would change our minds, and suspect many other skeptics would as well, about the legal standing of implicit prejudice work to the degree IAT advocates could show that: (1) even if we accept their meta-analytic estimates of the effect sizes for implicit prejudice, the data are fully consistent with modest estimates of the prevalence of prejudice in the populations thus far studied; (2) the predictive power of their new measures of prejudice has not been artificially inflated by publication bias or by outliers, or simply represents the influence of old-fashioned prejudice; and (3) most critical, their measures predict legally actionable discrimination in realistic work settings, including subtle acts of discrimination in subjective ratings of candidates. It would now be instructive if implicit prejudice advocates would indicate, ex ante, what evidence would induce them to modify or abandon their own causal and value claims.