FACTS DO MATTER: A REPLY TO BAGENSTOS

*Gregory Mitchell**
*Philip E. Tetlock***

I. INTRODUCTION

Before the ink had dried on our critical assessment of new empirical research on implicit bias,[1] Professor Samuel Bagenstos had critiqued our critique.[2] Unfortunately, Bagenstos resorted to an old rhetorical gambit: refute a caricature of an opponent's position rather than the real thing. If Bagenstos were to be believed, we took the position that the "the law should prohibit only discrimination that results from self-conscious, irrational animus"[3] and that "rational discrimination is not a proper target of antidiscrimination law."[4] Having attributed this extreme position to us, it is no wonder that Bagenstos concludes that our arguments will "resonate strongly" with those who believe "antidiscrimination law should aim solely at punishing individuals who act on irrational racist and sexist animus" but "will be far less relevant" to those who see discrimination as "a social problem that entrenches the subordinated and disadvantaged status of particular, socially salient groups."[5]

Bagenstos is not alone in his misreading of our article. Professors Lane, Kang, and Banaji recently dismissed our article as a "predictable political backlash" to the far-ranging policy implications of implicit bias research,[6] and Professors Benforado and Hanson likewise described our article as part of a "backlash" against the implications of this research

---

* Daniel Caplin 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 helpful comments on an earlier draft by John Monahan, George Rutherglen, Rip Verkeke, Larry Walker, and Amy Wax.

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

1. *See generally* Gregory Mitchell & Philip E. Tetlock, *Antidiscrimination Law and the Perils of Mindreading*, 67 OHIO ST. L.J. 1023 (2006).

2. *See generally* Samuel R. Bagenstos, *Implicit Bias, "Science," and Antidiscrimination Law*, 1 HARV. L. & POL'Y REV. 477 (2007). We sent a draft of this Article to Professor Bagenstos, but he never responded to our message seeking his input.

3. *Id.* at 480.

4. *Id.* at 488.

5. *Id.* at 490.

6. Kristin A. Lane, Jerry Kang & Mahzarin R. Banaji, *Implicit Social Cognition and Law*, 3 ANN. REV. L. & SOC. SCI. 427, 442 (2007); *see also id.* at 439 ("[A]ssuming that the scientific research continues along its current trajectory, implicit social cognition has the potential to influence the understanding of intent in all bodies of law.").

that supposedly threatens cherished but incorrect views about human nature.[7] We suspect that misunderstandings of this magnitude could have arisen only if we were singularly inept in articulating our own position or if our scholarly adversaries were singularly determined to discredit our position by portraying it as far more extreme than it is. Not surprisingly, we favor the second interpretation—and we present in this reply to our critics, particularly Professor Bagenstos, evidence to support our interpretation. In the process, we make clear why legal scholars of any prescriptive bent should care about the unresolved empirical questions swirling round the concept of implicit bias.

It is our view that the legal community is under no obligation to agree when a segment of the psychological research community labels the vast majority of the American population unconsciously prejudiced on the basis of millisecond reaction-time differentials on computerized tests.[8] It is our view that the legal community should require evidence that scores on these tests of "unconscious prejudice" map in replicable functional forms onto tendencies to discriminate in realistic settings and that proposed remedies actually work before making wholesale changes to antidiscrimination law and policy. And it is our view that scientifically-informed public policy requires critically scrutinizing knowledge claims, not passively acquiescing to the latest shifts in intellectual fashions in carefully selected sub-fields of the notoriously fragmented social sciences.

II. SETTING THE RECORD STRAIGHT

Professor Bagenstos does not pinpoint any passages in our article where we endorsed the normative positions that he ascribes to us.[9] Bypassing a tedious bottom-up textual approach to our article, Bagenstos adopts a top-down deductive strategy in which he argues that only scholars committed to a narrow, perpetrator-focused model of illegal discrimination could find criticisms such as ours relevant to antidiscrimination law; hence, we must hold that narrow normative view.[10]

---

7. Adam Benforado & Jon Hanson, *Legal Academic Backlash: The Response of Legal Theorists to Situationist Insights*, 57 EMORY L.J. 1087, 1135-43 (2008).

8. For a thorough discussion of Implicit Association Tests, see *infra* notes 48-53 and accompanying text.

9. He admits as much: "[Mitchell and Tetlock's] arguments fail as a scientific challenge to implicit bias research because they largely rest on a set of *unexpressed* normative disagreements with implicit bias scholars." Bagenstos, *supra* note 2, at 493 (emphasis added).

10. *See id.* at 488. Bagenstos attributes to us the view that "Alan Freedman [sic] called the 'perpetrator perspective.'" *Id.* at 488 (quoting Alan David Freeman, *Legitimizing Racial*

Before showing how Bagenstos misunderstands our criticisms and the implications of these criticisms for *any* theory of antidiscrimination law founded on the science of implicit bias, we pause to note that Bagenstos not only reads too much into our arguments—he ignores direct evidence that we do not hold the normative position that he ascribed to us. We acknowledged that "[a]rguably, under current law, disparate treatment claims can encompass acts of discrimination caused by conscious and unconscious influences."[11] We did not argue that rational discrimination should be legal and expressly warned against that misunderstanding of our arguments: "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*.'"[12] We said that "[i]t is difficult to overstate the legal significance of [the implicit bias] research if it correctly diagnoses the pervasiveness and potency of implicit prejudice and related discriminatory tendencies," acknowledging the many ways that valid and rigorous implicit bias work might be used to alter the law.[13] And we defined bias and prejudice expansively, to include psychological processes operating at an implicit level, while acknowledging that legal definitions might be narrower but *without* endorsing the view that legal definitions *should* be narrower.[14]

Discrimination Through Antidiscrimination Law: A Critical Review of Supreme Court Doctrine, 62 MINN. L. REV. 1049, 1052-54 (1978)).

11. Mitchell & Tetlock, *supra* note 1, at 1038 n.47. Lest we be misread again, we note that our use of the qualifier "arguably" here was simply a reflection of uncertainty within the law rather than any normative judgment on our part. *See, e.g.*, Samuel R. Bagenstos, *The Structural Turn and the Limits of Antidiscrimination Law*, 94 CAL. L. REV. 1, 8 (2006) ("There is some question whether existing antidiscrimination law even prohibits actions driven by unconscious bias.").

12. Mitchell & Tetlock, *supra* note 1, at 1085 (emphasis added). One of us has even conducted empirical research into the psychology of statistical discrimination and discussed the very good grounds that may exist for forbidding use of certain base rates, no matter how true they may be. *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, 863-69 (2000).

13. Mitchell & Tetlock, *supra* note 1, at 1054.

14. *See id.* at 1035-38. In advancing his argument that we are committed to an old-fashioned view of prejudice, Bagenstos also misrepresents, perhaps through misunderstanding, our point about divergent validity among implicit and explicit intergroup attitudes. Bagenstos asserts that we assume "that implicit bias reflects explicit prejudicial attitudes," and that we contend the IAT research lacks construct validity because the correlations between implicit and explicit attitudes are inconsistent. Bagenstos, *supra* note 2, at 482. Nowhere do we assume that implicit bias must be a reflection of explicit bias for the former to be valid. Our simple point was that IAT researchers have taken contradictory positions on whether implicit and explicit attitudes should converge or diverge, contradictions that demonstrate how preliminary our knowledge in this area is and reveal the lack of

For the record—again—we do not defend the view that rational or unintentional discrimination should be legal or that illegal discrimination must involve animus. Indeed, we do not reject public health approaches to discrimination, even of the kind that Bagenstos favors,[15] as they may have value depending on details of implementation. But we see no value in basing any model of discrimination—no matter how nobly intentioned—on flawed social science or basic research with no demonstrated external validity for real work settings. That was the overriding message of our earlier article.

We now turn to the specific arguments Bagenstos advances to try to show that our criticisms of implicit bias research have bite only if one endorses a narrow definition of illegal discrimination. We show that Bagenstos makes a number of errors as he tries to reconcile our criticisms with his preferred interpretations of the implicit bias research and his normative view on antidiscrimination law.

III. CAN WE FORMULATE EFFECTIVE REMEDIES WITHOUT UNDERSTANDING THE NATURE OF IMPLICIT BIAS?

Our original article highlighted a host of ongoing debates about the validity of the most popular measure of implicit bias: a speeded stimulus-classification task known as the Implicit Association Test ("IAT"). These debates revolve around what causes some people to be slower when pairing stimuli from minority groups with positively-valenced rather than negatively-valenced stimuli.[16] An "implicit-bias-as-

---

an underlying psychometric theory for implicit bias. *See* Mitchell & Tetlock, *supra* note 1, at 1060-65.

15. *See* Bagenstos, *supra* note 2, at 491.

16. The IAT is the most popular implicit measure of bias, and IAT research forms the basis for most of the new antidiscrimination law scholarship on implicit bias. *See, e.g.*, Tristin K. Green, *A Structural Approach as Antidiscrimination Mandate: Locating Employer Wrong*, 60 VAND. L. REV. 849, 854 (2007) ("Perhaps the most well-known recent work on the pervasiveness of unconscious bias is the Implicit Association Test ('IAT'), developed by social psychologists Anthony Greenwald and Mahzarin Banaji."). The IAT is a speeded binary-classification task in which test-takers are shown stimuli on a computer screen and asked to press a key on the right or left side of the keyboard to put the stimuli into one of two categories. In the case of the race IAT, for instance, half the stimuli are photos of either Black or White faces, and the other half of the stimuli are words that are either positive or negative in character. For half of the trials, the individual presses one key if the face is White or the word is positive and a different key if the face is Black or the word is negative. For the other half of the trials, the individual presses one key if the face is White or the word is negative and a different key if the face is Black or the word is positive. It is assumed that an individual who is implicitly biased against Blacks will be slower to respond in the latter task than the former task. The latencies of each response on a given trial are recorded (in milliseconds) and these latencies are then combined in a complex fashion to yield an overall score that reflects the implicit attitude. Scores are interpreted such that positive scores indicate an implicit preference for Whites over Blacks and negative scores indicate an implicit preference for Blacks

pure-prejudice" account treats a positive differential in reaction times between the minority-negative and minority-positive trials on the IAT as reflecting unalloyed unconscious negativity toward these groups.[17] Although not even the IAT creators embrace the view that the IAT is "process-pure" (that is, that it measures only the psychological process of interest),[18] some legal scholars seem to think that it is.[19] Contrary to the latter view, the research literature reveals a host of alternative explanations for, and possible influences on, this reaction-time differential, including lesser familiarity with minority groups, test anxiety or evaluation apprehension, guilt about discrimination, sympathy for disadvantaged groups, cultural knowledge, and a test-taker's cognitive dexterity in switching between mental tasks.[20] Professor Bagenstos reads our discussion of these alternative accounts of what influences the IAT scores as an argument that the law should not concern itself with implicit bias if it does not involve the "narrow sort of prejudice" that he wrongly asserts we believe can be the only proper source of regulation.[21] He then dismisses our whole discussion as irrelevant: "There is no reason why it should make a difference, from an antidiscrimination law and policy perspective, whether implicit bias reflects 'hostility' toward minorities or any of the alternative explanations Mitchell and Tetlock offer."[22]

Here readers can see precisely where our and Bagenstos's positions diverge. For Bagenstos, it is self-evident that the IAT directly taps into implicit bias: after all, people are apparently having a harder time recognizing and responding to pairings of black-good than to pairings of white-good—or an easier time responding to pairings of black-bad than

---

over Whites. *See generally* Kristin Lane et al., *Understanding and Using the Implicit Association Test: IV: What We Know (So Far) About the Method*, *in* IMPLICIT MEASURES OF ATTITUDES 59 (Bernd Wittenbrink & Norbert Schwarz eds., 2007) (describing IATs and discussing how biases are calculated based on the results of such tests).

17. *See* Lane et al., *supra* note 16, at 72.

18. Brian Nosek, one of the creators of the IAT, recently stated that "no measure is a process-pure assessment of a construct." Brian A. Nosek & Frederick L. Smyth, *A Mutlitrait-Multimethod Validation of the Implicit Association Test: Implicit and Explicit Attitudes Are Related but Distinct Constructs*, 54 EXPERIMENTAL PSYCHOL. 14, 25 (2007).

19. *E.g.*, Jerry Kang, *Trojan Horses of Race*, 118 HARV. L. REV. 1489, 1510 (2005) ("Tasks in the schema-consistent arrangement should be easier, and so it is for most of us. How much easier—as measured by the time differential between the two arrangements—provides a measure of implicit bias. The obvious confounds—such as overall speed of participant's reactions, right- or left-handedness, and familiarity with test stimuli—have been examined and shown not to undermine the IAT's validity.") (citations omitted).

20. *See* Mitchell & Tetlock, *supra* note 1, at 1072-93.

21. Bagenstos, *supra* note 2, at 483.

22. *Id.* at 485.

to those of white-bad.[23] The IAT must therefore be gauging discrimination in microcosm—and Bagenstos apparently sees no problem scaling up to American race relations at large. If around eighty percent of the American population fails the IAT, that surely means that around eighty percent of the population must harbor an unconscious propensity to discriminate against African-Americans. And Bagenstos seems perplexed why anyone would care about nit-picky psychological details about whether this unconscious propensity to discriminate is rooted in animus as opposed to figure-ground asymmetries or cognitive dexterity or realistic assessments of covariation between group membership and bad outcomes. If people are being victimized by prejudice, whether it takes the form of conscious or unconscious bias, why do Mitchell and Tetlock want to take us on a detour into a potentially endless debate over the precise mechanisms causing the victimization?

By contrast, our view is that the validity of psychological tests must be demonstrated, not simply stipulated. We are certain that the IAT does not tap into "100% pure prejudice," whether implicit or explicit, and we remain unconvinced about the policy implications of the IAT work for two mutually reinforcing reasons: (1) an extensive body of psychological evidence that shows that IAT scores are influenced by a wide range of processes that cannot plausibly be labeled precursors to discrimination, and (2) the lack of psychological evidence that IAT scores map in strong and replicable functional forms onto propensities to discriminate in the real world. Put as simply as we know how, the fact that eighty percent of Americans "fail" the IAT does not mean that eighty percent of Americans are unconsciously biased. From our standpoint, it is inappropriate to claim an empirical mandate for restructuring antidiscrimination law on the basis of a still-developing line of scientific inquiry with no record of applied success.

This discussion brings us to the centerpiece irony of our exchange with Bagenstos. The normative conclusions that Bagenstos draws about our supposed implicit value agenda seem to be driven, at least in part, by scientific misconceptions that Bagenstos holds about implicit bias research. Bagenstos's pivotal misconception is that the much-publicized IAT failure rate (often placed around eighty percent) must be telling us something profoundly important about the true level of unconscious

23. For instance, Bagenstos states that "[i]f whites *act in a bigoted way on the IAT* because they are afraid of being stereotyped as bigots . . . whites who make employment decisions could just as well act in a bigoted way in those decisions because they are afraid of being stereotyped as bigots." *Id.* (emphasis added). Thus, for Bagenstos, the IAT reaction times themselves reveal a bigotry regardless of the source of the "bigotry."

prejudice in America today. But the IAT failure rate tells us little or nothing of importance about unconscious prejudice if it turns out that IAT scores have little or no power to predict real-world discrimination, no matter how expansively defined.[24]

Moreover, even if the IAT had been shown reliably to predict discrimination, it is still important to understand the underlying mechanisms of implicit bias. Policy that seeks to change unwanted outcomes without an understanding of the causes of these unwanted outcomes is likely to be ineffective and may even make matters worse.[25] To invoke a few simple examples from our earlier article, if implicit bias flows from lack of familiarity with minority groups, then policies that promote intergroup contact might be most effective, whereas bias flowing from cultural knowledge might be more effectively attacked by media policies of the sort Kang advances (and we would still need to know a great deal about causal mechanisms to formulate an effective media policy).[26] It seems to us that the implicit bias case for affirmative action measures of the sort Kang and Banaji propose weakens to the extent that test anxiety or differences in cognitive dexterity contribute significantly to IAT scores.[27] And should we introduce measures that raise consciousness about possible implicit biases, as Kang and Banaji suggest we should by having employees take the IAT,[28] if these measures will create greater anxiety and guilt about race relations and lead to the very kind of stilted, uncomfortable interracial interactions that implicit bias scholars contend disadvantage minorities in the workplace?[29] Are these measures even necessary if small amounts of

---

24. We emphasized this point in our original article. *See* Mitchell & Tetlock, *supra* note 1, at 1033, 1100, 1102, 1107-15. For further discussion of this point, see Amy L. Wax, *The Discriminating Mind: Define It, Prove It*, 40 CONN. L. REV. 979, 984-85 (2008).

25. Some evidence suggests that this has been the case with the Americans with Disabilities Act, for instance. *See* Samuel R. Bagenstos, *Has the Americans with Disabilities Act Reduced Employment for People With Disabilities?*, 25 BERK. J. EMP. & LAB. L. 527, 562 (2004) ("[T]o achieve their goals disability rights advocates must honestly confront the ADA's record in all of its complexities. The record has been a mixed one, and the disability rights movement will do itself no favors if it ignores the negative, as well as the positive, aspects of that record.").

26. Kang, *supra* note 19, at 1549-53.

27. *See* Jerry Kang & Mahzarin R. Banaji, *Fair Measures: A Behavioral Realist Revision of "Affirmative Action,"* 94 CAL. L. REV. 1063, 1098-101 (2006).

28. *Id*. at 1090-92.

29. *See, e.g.*, *id.* at 1095 ("Although awkwardness might seem trivial, prior research confirms that awkwardness leads to worse interviews."). Professor Norton and colleagues empirically demonstrated that one's efforts to appear unprejudiced in interracial interactions may have perverse effects: "Ironically, those Whites who tried hardest to appear color-blind by avoiding the use of race [in a task where race was relevant to a discussion] were the individuals who appeared least friendly when interacting with Black partners." Michael I. Norton et al., *Color Blindness and Interracial Interaction: Playing the Political Correctness Game*, 17 PSYCHOL. SCI. 949, 952 (2006).

individuating information moderate both implicit and explicit bias?[30] In our view, a better understanding of the causes of implicit bias should precede the formation of public policy if the goal is to formulate effective policy.

Our concern was, and is, that so many legal scholars have accepted—seemingly at face value—claims of implicit bias researchers that they have discovered powerful psychological precursors of discrimination operating at an unconscious level.[31] Bagenstos, for instance, treats the terms "implicit bias" and "unconscious bias" as synonyms.[32] If one assumes that implicit bias operates beyond conscious control and that the unconscious cannot be motivated by pro-diversity values or situational goals, then it is easy to conclude that antidiscrimination law places too much emphasis on the regulation of conscious thought and intentional conduct.[33] But is it correct to assume that implicit biases operate beyond control or beyond the influence of conscious knowledge and that the unconscious is not goal driven?

Empirical research suggests not. Conrey and her colleagues found that the IAT measures both controlled psychological processes (which do not meet the criteria in psychology for being classified as unconscious processes) and automatic psychological processes (which typically do meet those criteria).[34] And Professor De Houwer found that conscious propositional knowledge can influence IAT responses, leading him to conclude that "one cannot simply assume that implicit, reaction-

---

30. *See* Thomas R. Cain & Lee Jussim, Individuating Information Reduces Implicit and Explicit Stereotyping 15 (Oct. 26, 2007) (unpublished manuscript, on file with the Hofstra Law Review) ("The main message of this study is that individuating information reduces implicit stereotyping of individuals as much as it reduces explicit stereotyping of individuals.").

31. *See, e.g.*, Anthony G. Greenwald & Linda Hamilton Krieger, *Implicit Bias: Scientific Foundations*, 94 CAL. L. REV. 945, 946 (2006) (describing "implicit bias" as "an aspect of the new science of unconscious mental processes that has substantial bearing on discrimination law"); *id.* at 951 ("The very existence of implicit bias poses a challenge to legal theory and practice, because discrimination doctrine is premised on the assumption that, barring insanity or mental incompetence, human actors are guided by their avowed (explicit) beliefs, attitudes, and intentions.").

32. *See* Bagenstos, *supra* note 2, at 477 ("In recent years, scholars of antidiscrimination law have increasingly focused on the problem of 'implicit' or 'unconscious' bias.").

33. *See* Christine Jolls & Cass R. Sunstein, *The Law of Implicit Bias*, 94 CAL. L. REV. 969, 980 (2006) ("The central focus of existing antidiscrimination law is on prohibiting consciously biased decisionmaking—a focus that has produced intense criticism from those interested in implicit bias.").

34. 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. Clearly, attempts to overcome these associations also contribute to performance on these tasks.").

time based measures are impervious to the effects of conscious propositional knowledge."[35] Other research has found that subjects can fake both established and novel attitudes on the IAT when instructed to do so or when placed into situations where faking an attitude would be beneficial.[36] And there is now considerable evidence that even unconscious processes are goal driven, with egalitarian goals moderating implicit biases.[37] In short, the research picture that is emerging is much

35. 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).

36. *See* Sandor Czellar, *Self-Presentational Effects in the Implicit Association Test*, 16 J. CONSUMER PSYCHOL. 92, 99 (2006) ("Although these results indicate that the IAT is susceptible to self-presentation and that it is possible to fake the IAT, additional research is needed to understand how exactly these effects arise."); Jan De Houwer et al., *Novel Attitudes Can Be Faked on the Implicit Association Test*, 43 J. EXPERIMENTAL SOC. PSYCHOL. 972, 976 (2007) ("Our results suggest that participants can fake having newly formed attitudes as measured by the IAT."); Klaus Fiedler & Matthias Bluemke, *Faking the IAT: Aided and Unaided Response Control on the Implicit Association Tests*, 27 BASIC & APPLIED SOC. PSYCHOL. 307, 314 (2005) ("The three experiments reported here provide a clear-cut answer to the question of whether controlled responding or faking on the IAT is possible. The answer is an unqualified Yes."). Professors Fiedler and Bluemke tested whether experts could reliably detect faked IAT scores so that these scores could be removed from datasets and found that the experts could not reliably identify the faked scores. Fiedler & Bluemke, *supra*, at 315 ("For two experienced experts, it was virtually impossible to identify IAT fakers among [twenty-four] respondents above chance. Thus, fakers not only managed to produce desired test outcomes but also to conceal their controlled responding effectively.").

37. *See, e.g.*, Luigi Castelli & Silvia Tomelleri, *Contextual Effects on Prejudiced Attitudes: When the Presence of Others Leads to More Egalitarian Responses*, 44 J. EXPERIMENTAL SOC. PSYCHOL. 679, 683-84 (2008) ("The actual presence of other individuals significantly modified spontaneous responses demonstrating that changes in the surrounding social environment were effective in producing a positive shift in automatic intergroup attitudes. . . . In the presence of other persons, egalitarian-related concepts were more easily accessed after the presentation of Black faces demonstrating that the group context triggered the goal of not being prejudiced and behaving fairly toward Blacks.") (citation omitted); 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, 845 (2002) ("[W]e expected that participants who reported high levels of internal motivation and low levels of external motivation and, thus, were theoretically highly autonomous would be the most effective at regulating expressions of race bias, even on difficult-to-control responses. Consistent with our expectations, these individuals responded with lower levels of implicit race bias than did all other participants."); Gordon B. Moskowitz et al., *The Implicit Volition Model: On the Preconscious Regulation of Temporarily Adopted Goals*, 36 ADVANCES EXPERIMENTAL SOC. PSYCHOL. 317, 396 (2004) ("People who had fairness goals triggered by contemplating such a prior experience (thinking about failures at being egalitarian) did not show activation of the stereotype. They compensated for the triggered goal by engaging in goal-compatible responses such as inhibiting the stereotype."); Kai Sassenberg & Gordon B. Moskowitz, *Don't Stereotype, Think Different! Overcoming Automatic Stereotype Activation by Mindset Priming*, 41 J. EXPERIMENTAL SOC. PSYCHOL. 506, 511-12 (2005) (describing research where situational cues to think creatively activated a "think different" mindset in which stereotypic associations were inhibited, not activated, or not retrieved); Stacey Sinclair et al., *Social Tuning of Automatic Racial Attitudes: The Role of Affiliative Motivation*, 89 J. PERSONALITY & SOC. PSYCHOL. 583, 590 (2005) ("The reported experiments provide clear support for the notion that automatic racial attitudes are subject to affiliative social tuning. Across two experiments, two

more complex than that painted by proponents of a strong implicit bias view who dismiss regulation aimed at the intentional level.[38] At most, implicit bias research reveals that some automatic psychological processes can contribute to variation in scores on implicit measures of bias, but the degree to which automatic versus controlled processes contribute to these scores and the degree to which individuals can exert control over implicit bias (consciously or unconsciously) remain unclear.

Furthermore, not all automatic psychological processes call for the same regulatory response:

> Some automatic associations or impulses will drive behavior because they are very strong. In other cases, the automatic components will drive behavior because the ability to control them is weak. The implications of the two situations are very different. For example, the causes, consequences, and cures for implicit prejudice stemming from strong automatic biases and weak self-regulatory processes are going to be very different. In some contexts, it may be that training someone to overcome their biases is far more simple than eliminating those biases. Indeed, among the recent spate of research showing the apparent ease of changing implicit attitudes, we suspect that, in many cases, it has been a matter of training people to overcome their biases rather than eliminating them.[39]

Simply labeling a process automatic or unconscious does not dictate how the law should respond.

There is yet another reason to care about the specific processes that contribute to scores on implicit measures of bias: because even the simplest psychological tests will implicate a number of psychological processes that may differ in form and function across individuals, every test presents the potential for contamination by measurement artifacts. Although Bagenstos works hard to explain all influences on IAT scores as consistent with an implicit bias perspective,[40] the creators of the IAT

---

measures of automatic prejudice, and three operationalizations of affiliative motivation, automatic prejudice shifted toward the ostensible attitudes of a social actor to the degree that individuals were motivated to get along with him or her. . . . This research adds to the growing literature demonstrating that automatic attitudes flexibly respond to social motives."). *See generally* Gregory Mitchell, *Second Thoughts*, 40 MCGEORGE L. REV. (forthcoming 2009) (manuscript on file with the Hofstra Law Review) (arguing that second thoughts can act as important checks on judgments, decisions, and behaviors, with regard to interpersonal biases).

38. *See, e.g.*, Kang & Banaji, *supra* note 27, at 1079 ("[A]n explicit ex ante exhortation not to be intentionally unfair will do little to counter implicit cognitive processes, which take place outside our awareness yet influence our behavior.").

39. Jeffrey W. Sherman, *On Building a Better Process Model: It's Not Only How Many, but Which Ones and by Which Means?*, 17 PSYCHOL. INQUIRY 173, 177 (2006) (citations omitted).

40. As one example of Bagenstos's efforts to assimilate all evidence to the view that IAT scores signify socially-harmful implicit biases, consider his response to evidence that figure-ground

asymmetries (or the differential familiarity of IAT stimuli) inflate IAT scores by systematically slowing and speeding reaction times. *See* Mitchell & Tetlock, *supra* note 1, at 1072-79. Bagenstos writes that:

> Racial unfamiliarity ('figure-ground asymmetry' in Mitchell and Tetlock's technical term) exists in the employment setting as well as in the laboratory setting. To the extent that unfamiliarity with members of minority groups in certain work settings leads their white counterparts to engage (even unconsciously) in aversive behavior, that aversion will have the same negative effects on minorities' opportunities as will aversion motivated by hostility.

Bagenstos, *supra* note 2, at 485. Here Bagenstos either misreads the IAT research or makes the kind of erroneous inferential leap that we feared would happen with legal scholars' uses of the IAT work: evidence that figure-ground asymmetries inflate scores on the IAT by affecting millisecond reaction times on this computer-based test provide no evidence whatsoever that figure-ground asymmetries in the workplace lead to aversive behavior of any kind toward minorities. A score showing "bias" on the IAT is not empirical proof that this test-taker will respond to figure-ground asymmetries in a work environment the way he or she responded to stimulus asymmetries detected in milliseconds on a speeded, computer-based test, much less that he or she will engage in discriminatory behaviors in interpersonal relations. *See* Mitchell & Tetlock, *supra* note 1, at 1065-79, 1100-05. Bagenstos provides no logical, theoretical, or empirical basis for equating figure-ground asymmetry effects on the IAT with figure-ground asymmetry effects on any kind of behavior in a work setting.

Moreover, we addressed evidence on the question of whether IAT scores predict discrimination of any kind in interracial interactions and noted the weak and mixed results of the few studies that have sought to study this question, including mentioning one study that found that persons scoring as more biased on the IAT were rated as having better interactions with minorities than persons showing less bias on the IAT. *See id.* at 1065-72, 1100-05. In response to our mention of this particular study, Bagenstos writes that the study presents no problems for implicit bias research, because "[a] key point of implicit bias, to proponents of the theory, is that it is invisible. That minorities—many of whom may themselves share those biases—cannot detect implicit biases in others is not in any way a refutation of the theory." Bagenstos, *supra* note 2, at 483 (citations omitted). Again, Bagenstos misreads the implicit bias work or misapplies it as he tries to defend the implicit bias view: While implicit bias researchers have claimed that the psychological construct of "implicit bias" might be "invisible" to the actor and perhaps the target as well at times, they have not claimed that the behavioral consequences of implicit bias cannot be detected. (Indeed, if they had so argued, then it would be impossible to claim a scientific basis for the claim that implicit bias leads to discrimination because the hypothesis would be untestable.) In fact, to support the view that the IAT predicts discrimination, proponents of the strong implicit bias view often cite a study by Professors McConnell and Leibold, in which both observers and persons interacting with experimental subjects supposedly rated those subjects with higher IAT scores as more biased in their interracial interactions. Allen R. McConnell & Jill M. Leibold, *Relations Among the Implicit Association Test, Discriminatory Behavior, and Explicit Measures of Racial Attitudes*, 37 J. EXPERIMENTAL SOC. PSYCHOL. 435, 440 (2001). Thus, implicit bias researchers are quite willing to take a position contrary to the invisibility position advanced by Bagenstos when the data turn out differently than in the earlier study we mentioned. (Incidentally, our reanalysis of the McConnell and Leibold dataset reached very different conclusions from those reported by McConnell and Leibold, with our reanalysis finding that higher IAT scores were actually correlated with more positive interracial interactions as judged by observers.) Hart Blanton et al., *Strong Claims and Weak Evidence: Reassessing the Predictive Validity of the IAT*, 94 J. APPLIED PSYCHOL. (forthcoming 2009) (manuscript at 18, on file with the Hofstra Law Review). So we are left scratching our heads: Does Bagenstos believe that "bias" detected by the IAT will lead to "aversive behavior" toward minorities, or does he believe that "bias" detected by the IAT will always remain invisible except on the IAT, so invisible that even persons showing a strong bias on the IAT will treat minorities better than—or no worse than—persons showing no bias on the IAT? If he takes the first position that implicit bias harms minorities but claims that it does so in ways that cannot be

do not agree that all influences on IAT scores reflect intergroup bias. In 2003, five years after the IAT had been introduced and a tremendous amount of IAT data had been gathered, Greenwald, Banaji, and Nosek acknowledged that performance on the IAT is contaminated by what might be called general processing speed (in simple terms, how fast one can perform mental tasks such as reacting to stimuli flashed on a computer screen for purposes of categorizing these stimuli), such that persons with slower processing speed (for example, many older persons and persons with less working memory) will have higher "bias" scores on the IAT regardless of whether these persons associate the categories of minorities and women with more negative (or less positive) concepts than the category of White males. In other words, the pre-2003 version of the IAT artificially inflated the bias scores of some groups of persons simply because of how the test was administered and scored. Due to this measurement artifact, Greenwald and colleagues rejected the original method for scoring the IAT in favor of a new method that is supposed to remove this artifact.[41] Had researchers not scrutinized the sources of IAT scores, then persons with relatively slower processing speeds would continue to be labeled more biased than is justified due to this measurement artifact.[42]

Our point is not that legal scholars should do empirical research on measurement artifacts in the IAT. Rather, legal scholars need to be informed that data based on pre-2003 versions of the IAT

---

detected by the interacting parties, researchers, or courts, then his claim cannot be falsified. If he takes the second position, then there are no adverse behavioral effects of the bias detected on the IAT that could support a new regulatory regime.

41. Anthony G. Greenwald et al., *Understanding and Using the Implicit Association Test: I. An Improved Scoring Algorithm*, 85 J. PERSONALITY & SOC. PSYCHOL. 197, 212 (2003) ("The present findings call strongly for replacing the IAT's conventional scoring procedure."); *id.* at 214 ("[T]he improved algorithm offers a gain in construct purity. That is, the improved algorithm, compared with the conventional scoring procedure, is less contaminated by extraneous variables. One such contaminant is the conventional IAT measure's production of spuriously extreme IAT scores for slow responders . . . . The new algorithm almost completely eliminates this artifact . . . ."); *id.* at 214-15 ("Resistance to the response-speed artifact should be useful in studies that compare IAT scores for groups, such as children versus adults, that differ in speed of responding. The new algorithm likewise should provide more valid correlations of IAT measures with individual difference measures, such as age or working memory capacity, that correlate with response speed."). Whether the new method for scoring the IAT appropriately controls for processing speed artifacts remains a point of debate. *See, e.g.*, Hart Blanton et al., *Decoding the Implicit Association Test: Implications for Criterion Prediction*, 42 J. EXPERIMENTAL SOC. PSYCHOL. 192, 209-10 (2006) (discussing problems with alternative approaches to scoring the IAT).

42. Greenwald and colleagues acknowledge one other measurement artifact: prior experience taking the IAT tends to lead to lower IAT scores. The new algorithm "affords some protection" against this artifact but does not completely eliminate it. Greenwald et al., *supra* note 41, at 215. Thus, if one wants to become "less implicitly biased" (if IAT scores translate directly into implicit bias, as Bagenstos suggests), then one simply needs to practice taking the IAT.

unquestionably contain artificially inflated estimates of bias for some portion of the subject population and that other artifacts may exist given the very young age of the IAT and its undeveloped psychometric foundation.[43] Otherwise, bad advice may be given. For instance, before the article on the IAT's processing-speed artifact had been published, Professor Ayres suggested that IAT scores could be used "as a criterion for hiring both governmental and nongovernmental actors."[44] Because of the processing-speed artifact, however, an employer that followed Ayres' suggestion would probably have excluded from hiring a number of persons over the age of forty or fifty or anyone with lower working memory capacity (which tends to be correlated with lower IQ and not just age), regardless of those persons' "true" levels of implicit bias.[45]

---

43. Professor Borsboom describes the psychometrics underlying the IAT:

> First, the original paper [publishing the IAT procedure] puts forward no psychometric model for the dynamics underlying the test whatsoever. Second, even though the test is described as a measure of individual differences, the main evidence for its validity is a set of mean differences over experimental conditions, and no formal model explicating the link between these two domains is offered. Third, . . . the IAT is a popular measurement procedure despite these points. Fourth, it took no less than eight years for a detailed psychometric modeling analysis of the proposed measure to see the light; and that analysis suggests that the scoring procedures used are actually quite problematic, because the various possible psychometric models on which they could be predicated are not supported by the data.

Denny Borsboom, *The Attack of the Psychometricians*, 71 PSYCHOMETRIKA 425, 430 (2006) (citation omitted). And Borsboom does not even mention the low test-retest reliability coefficient of.56 associated with the IAT, which means that one's IAT score may vary widely over repeated testing sessions. *See* Brian A. Nosek et al., *The Implicit Association Test at Age 7: A Methodological and Conceptual Review*, *in* SOCIAL PSYCHOLOGY AND THE UNCONSCIOUS: THE AUTOMATICITY OF HIGHER MENTAL PROCESSES 265, 274 (John A. Bargh ed., 2007). We suspect that Bagenstos and other implicit bias proponents would roundly reject the use of an aptitude test with such a low reliability coefficient as part of an application or promotion process. *Cf.* REBECCA M. WARNER, APPLIED STATISTICS: FROM BIVARIATE THROUGH MULTIVARIATE TECHNIQUES 831-32 (2008) ("[W]hen scores for an individual are used to make decisions that have important consequences (such as medical diagnosis or placement in a special class for children with low academic ability), a reliability of r =.90 is the bare minimum and r =.95 would be a desirable standard for reliability (because even very small measurement errors could result in mistakes that could be quite harmful to the individuals concerned).").

44. IAN AYRES, PERVASIVE PREJUDICE? UNCONVENTIONAL EVIDENCE OF RACE AND GENDER DISCRIMINATION 424 (2001). Bagenstos expresses reservations about this suggestion. *See* Bagenstos, *supra* note 2, at 479, 481.

45. Bagenstos describes our raising the prospect of such false-positive accusations as showing "intense concern" for those who may be falsely labeled implicit racists or implicit sexists. Bagenstos, *supra* note 2, at 489. We fail to see how pointing out this fact shows "intense" concern for those falsely accused of implicit bias, much less how it reveals where we stand on the balancing of Type I errors (false negatives) and Type II errors (false positives). And we expressly stated that deciding whether any particular level of false accusations is excessive is a matter of political values rather than scientific fact. *See* Mitchell & Tetlock, *supra* note 1, at 1101. Absent some argument that Bagenstos does not make—perhaps Bagenstos feels that we must endure some number of false positives to prevent some number of true positives from being in the workplace—we see no value in

IV. SHOULD WE DEFER TO IMPLICIT BIAS RESEARCHERS' UNSCIENTIFIC JUDGMENTS ABOUT THE MEANING OF OBSERVED BEHAVIORS?

Professor Bagenstos's greatest irritation with our article seems to stem from our discussion of whether the IAT simply measures people's observations of objective conditions in the world.[46] Contrary to Bagenstos's suggestion,[47] we did not argue that discrimination flowing from rational observations of existing conditions in society should be legally immune, and we do not hold that view. Rather, we mentioned the possibility that the IAT measures tacit awareness of existing inequalities to illustrate the value-laden terms chosen by psychologists to describe data that are open to less loaded interpretations and to warn against placing too much weight on the tendentious labels chosen to characterize these data.

In our opinion, we have collectively entered the theater of the absurd if social psychologists can create a test that may measure whether people recognize inequalities in society and, when the test reveals that people do recognize indisputable inequalities, declare these test-takers to be implicitly biased against the have-nots in this society—*with no evidence that particular scores on this test predict any kind of behavior by particular test-takers toward the have-nots, whether that behavior be subtle or unsubtle, in the workplace, social life, or the voting booth*. Rather than question whether this conclusion follows, legal scholars and psychologists embrace this conclusion and argue in law reviews and courts that the results of this test prove that we have an epidemic of unconscious bias in our society.[48] To make the point crystal clear that we

---

excluding unbiased persons from the workplace through an erroneous test result. But again, that was not our point—we were just reporting this fact because it would seem to be a potentially relevant, but neglected, point for policy-making purposes.

46. *See* Bagenstos, *supra* note 2, at 486-88.

47. *Id.* at 488.

48. *See, e.g.*, Brief Amicus Curiae of the American Psychological Ass'n in Support of Respondents at 8, Grutter v. Bollinger, 539 U.S. 306 (2003) (Nos. 02-241 & 02-516) ("The results of the research on associative processes conclusively demonstrate that unconscious stereotyping and prejudice, including race stereotyping, is widespread. To take but one example, test respondents consistently make more ready associations between, on the one hand, faces of African Americans and words having negative concepts (*e.g.*, bomb, devil, awful) and, on the other, faces of Whites and positive concepts (*e.g.*[,] peace, joy, love). Hundreds of thousands of individuals have taken the test producing similar results.") (citation omitted); Bagenstos, *supra* note 11, at 5 ("Social psychologists have demonstrated that implicit biases against women and racial minorities remain widespread."); Kang & Banaji, *supra* note 27, at 1072 ("[B]y a conservative estimate, around ninety percent of Americans (and others in the western world), mentally associate negative concepts with the social group '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

are attacking the logic of reading the IAT as a direct pipeline to implicit bias with behavioral implications, and are *not* arguing that actual discrimination flowing from rational beliefs should be legal, imagine that the IAT is not a speeded stimulus-classification test taken on the computer, but rather a paper-and-pencil test that asks the test-taker whether, in America today, African-Americans as a group fare better or worse than European-Americans as a group. If the test-taker answers that African-Americans fare worse, then we score the test-taker as implicitly biased. If the test-taker asks on what basis we reach that conclusion, then we reply that: (a) our test is an implicit measure of bias because its purpose—to assess racial bias—was surely not obvious and hence the test-taker had no motive to fake the answer;[49] and (b) the test-taker's response reflects positive and negative mental associations with the categories "African-American" and "European-American."[50] If the label "implicitly biased" seems absurd to apply here, then one appreciates our point that the inventors of the IAT—who, from its inception, labeled the IAT a measure of "implicit prejudice" despite any evidence that it predicted any kind of discriminatory behavior[51]—have embraced an expansive view of implicit prejudice.

One may say that the inventors are entitled to claim whatever they wish about what their test measures. But the scientific community is

---

they associate female to family."); *see also* Greenwald & Krieger, *supra* note 31, at 955 ("Since 1998, IAT measures of implicit attitudes have been available on the Internet for self-administered demonstration use. These web-accessible demonstrations, which allow users to interactively experience the IAT, have accumulated sufficient data to allow researchers to draw conclusions about the pervasiveness of implicit and explicit biases.") (citation omitted).

49. Our imaginary version of the IAT thus continues to meet the IAT creators' definition of an "implicit" measure: "The IAT is an implicit measure because it infers group-valence and group-trait associations from performances that are influenced by those associations in a manner that is not discerned by respondents." Greenwald & Krieger, *supra* note 31, at 952. That a few respondents might guess our purpose will not require us to divest our instrument of the "implicit measure" label under the IAT creators' standards. *See id.* at 952 n.23.

50. *Cf.* Lane, Kang & Banaji, *supra* note 6, at 435 ("The culture and the person are intricately intertwined, and it would be a mistake to assume that signals of ingroup-favoring attitudes, because their content may have its origins in culturally shared knowledge, are hence not a part of the individual.").

51. *See* Anthony G. Greenwald et al., *Measuring Individual Differences in Implicit Cognition: The Implicit Association Test*, 74 J. PERSONALITY & SOC. PSYCHOL. 1464, 1475 (1998) (describing their findings as "discouraging in indicating the pervasiveness of unconscious forms of prejudice"). The preceding paper was the first publication of the IAT method and IAT data. *See also* Mahzarin R. Banaji, *Implicit Attitudes Can Be Measured*, *in* THE NATURE OF REMEMBERING: ESSAYS IN HONOR OF ROBERT G. CROWDER 117, 138-39 (Henry L. Roediger III et al. eds., 2001) ("[T]he IAT roughly detects repeated cultural pairing (moderated through individual experience) of Black + bad and White + good, most clearly among non-Black inhabitants of the United States. . . . Implicit attitudes, as I see it, reflect traces of experiences within a culture that have become so integral a part of the individual's own mental and social makeup that it is artificial, if not patently odd, to separate such attitudes into 'culture' versus 'self' parts.").

entitled to scrutinize these claims—and the legal and policy communities are entitled to be skeptical when labels linked to test results are being bandied about in debates over how racist American society is, or the tests are being used to give the general public feedback about its supposed unconscious biases and the strength of these biases. Our fear is that the public and legal scholars are deferring to social scientists in their choice of terms because they believe these terms to possess an epistemic authority that they do not. Recall that the output of the IAT is relative reaction times in classifying minority and majority stimuli paired with words putatively having positive or negative connotations. If you take the race IAT on the web and have slower reaction times when pairing minority stimuli with positive things, for example, you will not be told your reaction times but instead will be told that you have a "slight automatic preference," a "moderate automatic preference," or a "strong automatic preference" for European-Americans relative to African-Americans, depending on how much faster you acted when pairing majority stimuli with positive things relative to minority stimuli with positive things.[52] Perhaps the term "response bias" would be empirically justified by these data (assuming any reaction-time differential was not a measurement artifact), but to call millisecond reaction-time data on a stimulus-classification test evidence of an "automatic preference" implies that whatever caused the reaction-time differential will also cause one automatically to prefer Whites over Blacks in other contexts, even though the IAT does not measure preferential judgments or choices at a conscious or unconscious level (unless one embraces the extreme reductionist view that mental associations of group categories with valence terms *are* preferences). In light of the many alternative accounts of what the IAT might measure, there is no scientific justification for choosing the "automatic preference" account until the alternatives are eliminated, and there is no scientific justification for implying that "automatic preferences" as measured by the IAT generalize until ranges of IAT scores are empirically connected to preferences in other contexts. The personal and social implications of an unconscious "preference" for Whites over Blacks may be very troubling, but there is presently no scientific basis for reading IAT scores to signify such preferences.[53]

---

52. One may take the race IAT at Project Implicit, https://implicit.harvard.edu/implicit/demo/selectatest.html (last visited July 21, 2009).

53. And as we discuss in Part V, there is little evidence that the IAT correlates with discriminatory behavior (or group preferences) in the aggregate and no evidence that the IAT can make reliable distinctions between the behavior of individuals who score differently on the IAT. In any event, implicit bias researchers have needed nothing more than IAT scores to pronounce the great majority of Americans implicitly biased because an IAT score alone is treated as an unproblematic measure of relative group preferences.

Furthermore, the conversion of continuous reaction-time data from the IAT into qualitative categories, with these qualitative categories then being used to characterize the pervasiveness and strength of implicit biases within the United States population,[54] lacks scientific justification. The conversion of quantitative reaction-time scores into value-laden qualitative categories is not grounded in any theory about the functions relating IAT scores to the underlying latent construct of implicit bias, be it a linear or geometric function of some sort. Nor is it grounded in any empirical data about the behavioral implications of different ranges of IAT scores (for example, data showing that IAT scores in a certain range correlate consistently with opposition to affirmative action or negative reactions to Blacks or that people populating the same ranges of IAT scores show equivalent levels of bias on other measures of implicit bias). The conversion to qualitative categories reflects nothing more than the IAT architects' adaptation of a rule of thumb used by social scientists to compare effect sizes across studies.[55]

The originator of this rule of thumb, Jacob Cohen, did not propose that the convention be used to define the extremity of a score in a distribution and certainly not to convert continuous scores into qualitative categories for purposes of giving descriptive feedback to test-takers or population-level diagnostics. Cohen intended his rule of thumb simply as a shorthand way for researchers to compare the size of group means across studies, and he made clear that his rule of thumb was not based on empirical data about the social or practical meaning of different effect sizes:

> [T]hese proposed conventions were set forth throughout with much diffidence, qualifications, and invitations not to employ them if possible. The values chosen had no more reliable a basis than my own intuition. They were offered as conventions because they were needed in a research climate characterized by a neglect of attention to issues of magnitude. The [effect size] measures and conventions have been

54. *See, e.g.*, Greenwald & Krieger, *supra* note 31, at 955-56 (summarizing a large dataset of IAT scores collected via the web and characterizing this dataset as showing pervasive implicit biases in favor of advantaged groups and characterizing these biases in terms of strength); Robert G. Schwemm, *Why Do Landlords Still Discriminate (and What Can Be Done About It)?*, 40 J. MARSHALL L. REV. 455, 500-01 (2007) ("The Race IAT has been administered to tens of thousands of people in the United States, over eighty percent of whom exhibit an implicit pro-white bias. These results show that our unconscious racial attitudes 'may be utterly incompatible with our stated conscious values.' For example, even people who profess strongly egalitarian views often exhibit pro-white implicit biases." (citations omitted) (quoting MALCOLM GLADWELL, BLINK: THE POWER OF THINKING WITHOUT THINKING 85 (2005))).

55. *See* Greenwald et al., *supra* note 41, at 199 n.3 ("The *slight medium,* and *strong* labels corresponded to results meeting the conventional criteria for small, medium, and large effect sizes of Cohen's (1977) *d* measure.").

> successful, widely adopted not only for power analysis, but more widely, for example, in [effect size] surveys and in meta-analysis. But there are difficulties and much room for misunderstanding.[56]

The mischief that may come from endowing the classification scheme chosen by the IAT researchers with social meaning can be illustrated by examining the distribution of implicit bias as it existed in 2002, before the IAT architects acknowledged that a processing-speed artifact inflated bias scores, and the distribution in 2007, after changes were made to try to remove that artifact. In 2002, forty-eight percent of the people who took the race IAT were classified as showing a strong automatic preference for Whites over Blacks.[57] In 2007, twenty-seven percent of the people who took the IAT were classified as showing a strong automatic preference for Whites over Blacks.[58] No transformative legal or social event occurred within this five-year time period to reduce implicit bias by twenty-one percent; the IAT architects simply changed the way they defined bias on the test and, in connection with this change, employed a new variant of Cohen's effect-size rule of thumb (in other words, no research linking implicit attitudes to behavior drove the result—an analogy to Cohen's effect size convention and nothing more is wholly responsible for the categorical definitions).[59] The result was a dramatic change in the distribution of implicit bias in the United States, at least within the IAT creators' classification scheme (the table below provides a full breakdown of how percentages in the various qualitative categories changed from 2002 to 2007).[60]

Category	2002	2007
Strong Automatic Preference for Whites Over Blacks	48%	27%

56. JACOB COHEN, STATISTICAL POWER ANALYSIS FOR THE BEHAVIORAL SCIENCES 532 (2d ed. 1988).

57. Hart Blanton & James Jaccard, *Unconscious Racism: A Concept in Pursuit of a Measure*, 34 ANN. REV. SOC. 277, 291 tbl.1 (2008).

58. *Id.*

59. *Id.* at 291.

60. Mischief may occur at the individual level as well. First, because continuous scores are converted into categories using hard cut scores, one may be told one has a moderate automatic preference for Whites over Blacks when one's reaction times are much more similar to the average reaction time for those in the slight automatic preference category than the moderate category. Second, placing people into slight, moderate, and strong automatic preference categories based on relative position within a distribution can produce very misleading feedback. Thus, a person whose objective reaction times exhibit less of a differential than another person's may be categorized as more biased if the distribution containing the first person is tighter (in terms of the standard deviation) than the distribution containing the second person (assuming the same distribution means).

Moderate Automatic Preference for Whites Over Blacks	13%	27%
Slight Automatic Preference for Whites Over Blacks	12%	16%
No Preference	12%	17%
Slight Automatic Preference for Blacks Over Whites	6%	6%
Moderate Automatic Preference for Blacks Over Whites	4%	4%
Strong Automatic Preference for Blacks Over Whites	6%	2%

Table 1: Distribution of Race IAT Scores by Categories[61]

Imagine that, five years from now, the inventors of the IAT again decide to change their criteria for converting reaction times into qualitative categories, such that implicit bias surges back to pre-2007 levels or even higher if they should choose to make it go higher through definitional fiat. Should courts and legislatures defer to these social psychologists' judgments about what constitutes slight, moderate, and strong bias? Should courts and legislatures, as well as legal scholars relying on implicit bias work, look for empirical data to justify the conversions, or at least some theory about how IAT scores relate to the underlying psychological construct that is supposedly being measured?[62]

The upshot of many of Bagenstos's arguments is that we should defer to the implicit bias researchers' interpretations of the IAT data and to their definitional decisions, regardless of the scientific foundation for those interpretations and decisions.[63] Anyone who dissents must do so because he or she rejects a progressive antidiscrimination law agenda. Of course, that conclusion does not follow. A strategy in which one stops investigating the facts once the facts support preferred value

61. Adapted from Blanton & Jaccard, *supra* note 57, at 291 tbl.1.

62. Lest we be misunderstood again, we do not contend that the IAT could never meet some scientific test for validation as a psychometrically reliable measure of an individual's group bias or automatic preferences. It just has not done so yet.

63. As yet another example, Bagenstos endorses the implicit bias researchers' view that implicit bias is a unique psychological construct distinct in some way from explicit bias: "Scholars who advocate using the law [to respond to implicit bias] have strongly urged that implicit biases are meaningfully distinct from explicit attitudes." Bagenstos, *supra* note 2, at 482. Bagenstos is correct that legal advocates of the strong implicit bias view urge that implicit and explicit bias are distinct, but, from an evidentiary standpoint, the research does not point to such a simple conclusion. As we discussed previously, different implicit measures of bias do not converge as one might expect if they are tapping into the same latent constructs, and implicit and explicit measures of bias converge and diverge rather unpredictably and certainly to a greater extent than a strong "implicit bias is distinct from explicit bias" view can support. *See* Mitchell & Tetlock, *supra* note 1, at 1059-65. Our point in noting these open questions was not to argue that implicit and explicit bias must be one and the same or the law should look the other way. Our point is that there is not yet a sound scientific case for the "dissociation" view that Bagenstos endorses and which is so central to the argument that implicit bias calls for radical new forms of legal regulation. *See supra* note 14 (discussing how IAT researchers have taken contradictory positions on the convergence and divergence of implicit and explicit attitudes).

positions is not a wise strategy for scientific research, litigation, or policymaking. If the validity of prescriptive arguments based on implicit bias research depends on acceptance of the loaded data labels arbitrarily chosen by implicit bias researchers, then perverse effects are likely to follow from policy based on these prescriptive arguments, and these arguments are unlikely to persuade persons not already convinced of the need for new approaches to discrimination and societal inequality.

V. WHAT ROLE DOES SCIENCE REALLY PLAY IN BAGENSTOS'S ARGUMENT?

Professor Bagenstos ends his critique of our article with admirable candor about his own normative agenda and reasons for disputing the claims in our article. This candor reveals that, ultimately, Bagenstos's primary concern is that our scientific concerns will be mistaken as a refutation of the normative agenda that has come to be associated with the implicit bias research.

In closing his article, Bagenstos concedes that some of our challenges "call for caution and further study before taking the more extreme steps urged by some scholars who write about implicit bias," but he insists that our challenges "do not undermine the strong reasons to believe that implicit bias contributes to a serious social problem that denies opportunities to blacks and other minorities in America."[64] What are these "strong reasons" to believe implicit bias causes social harms?

We cannot say for sure, because Bagenstos does not directly state those reasons, but one possibility suggested elsewhere in his paper is that Bagenstos believes subtle acts of discrimination caused by implicit biases accumulate to the disadvantage of minorities.[65] Under this view, the dearth of present evidence showing that implicit measures of bias predict discriminatory behaviors at an observable level is not harmful to normative claims based on implicit bias.[66] A few points should be emphasized with respect to that view.

First, relatively few published studies have found any correlation between implicit bias and discriminatory behavior of any kind, subtle or unsubtle. Second, a meta-analysis by the creators of the IAT found an

64. Bagenstos, *supra* note 2, at 490.

65. *See id.* at 481 n.23 ("As workplaces evolve into more flexible, less hierarchical organizations, there is reason to believe that micro-level discriminatory behaviors will increasingly aggregate to cause denials of opportunities, but the 'reason to believe' is not yet hard proof.") (citation omitted); *see also id.* at 485 ("Whatever it reflects, implicit bias can result in behaviors and evaluations that limit the opportunities of minority group members.").

66. *See id.* at 481 n.23 ("[T]he link between implicit bias and discriminatory behavior that actually denies opportunities needs further development.").

average correlation of $r = 0.24$ between IAT scores and behaviors expansively defined to be indicators of racial stereotyping or racial prejudice (for example, the meta-analysis includes studies in which neurological responses to minority stimuli were the criterion behaviors).[67] If we accepted this estimate as reliable, the psychometric fact that IAT scores have low positive correlations with behavior expansively defined as discrimination would guarantee that many individuals labeled implicitly biased by the IAT will not exhibit discriminatory behavior of any kind. Third, to our knowledge no published study examines implicit bias and behavior toward non-strangers. While many important interactions will be between strangers (for example, job interviews, new client contacts), many important interactions also involve non-strangers, and one cannot reasonably assume that implicit biases toward broad social groups will automatically trump individualized associations with specific members of these groups. Fourth, studies on the IAT-behavior linkage fail to find a robust relationship between IAT scores and discriminatory behavior: the findings of such studies appear to be sensitive to the influence of outliers, and adding IAT scores to predictive models reduces prediction errors by only very small amounts.[68] Thus, while it is possible that evidence will develop supporting Bagenstos's cumulative disadvantage hypothesis, Bagenstos correctly surmises that no hard proof yet exists to support it.[69]

Another possibility is that Bagenstos feels that data connecting specific acts of discrimination to specific biased persons or organizations is unnecessary because he wants to reframe the discrimination problem as a matter of public health rather than individual fault or liability:

> Science does not defeat the implicit bias law-reform program, but science does not establish the case for that program, either. That program depends on a normative judgment that discrimination is not about fault but about a social problem—a normative judgment that is deeply contested among judges and policymakers. Unfortunately, arguments about implicit bias too often place heavy rhetorical emphasis on appeals to "science," perhaps at the expense of

---

67. *See* Anthony G. Greenwald et al., *Understanding and Using the Implicit Association Test: III. Meta-Analysis of Predictive Validity*, 96 J. PERSONALITY & SOC. PSYCHOL. (forthcoming 2009) (manuscript at 72, on file with the Hofstra Law Review). For IAT measures of sex bias and other intergroup biases, the correlations between IAT scores and behavioral indicators of bias were lower. *See id*. For discussion on prior versions of this meta-analysis, see Mahzarin R. Banaji et al., *No Place for Nostalgia in Science: A Response to Arkes and Tetlock*, 15 PSYCHOL. INQUIRY 279, 282 (2004) and Greenwald & Krieger, *supra* note 31, at 960; Kang & Banaji, *supra* note 27, at 1071-72.

68. *See* Blanton et al., *supra* note 40, at 17, 22-23, 30.

69. *See supra* note 40.

> developing the normative arguments for applying antidiscrimination law to that important phenomenon.[70]

Implicit bias in this view becomes a placeholder for the innumerable subtle ways society favors historically advantaged groups over historically disadvantaged groups. This view can be seen as a world view that operates at such a level of generality or abstraction, and permits so many defensive moves when threatened, that it becomes difficult to refute using ordinary means of empirical testing.[71]

Of course, one could endorse a move away from an actor-focused model and toward a public health model of antidiscrimination law without the prop of social scientific research on implicit bias—indeed, Bagenstos has done so in the context of disability law[72]—but to an advocate of the public health model, the allure of the implicit bias research is obvious: if this research demonstrates that discrimination occurs at an unconscious level, then it supports the view that the old, perpetrator-focused model cannot adequately address modern discrimination. As we have pointed out, however, both controlled and automatic psychological processes contribute to implicitly biased responses on the IAT and the term "unconscious" is not synonymous with "unmotivated." More generally, the jury remains out on exactly what psychological processes give rise to "biased" responses on implicit measures of bias and on how controllable these processes may be. Thus, even at the most basic level, it is not clear that the science of implicit bias supports the reform agenda that Bagenstos advocates.

In other words, Bagenstos's explicit agenda is to push antidiscrimination law and theory toward a public-health model, and, to the extent the new implicit bias work supports his agenda, he is happy to

---

70. Bagenstos, *supra* note 2, at 491.

71. Elsewhere we discuss the two polar positions within antidiscrimination theory that many scholars are drawn toward. While some scholars advocate a statist-interventionist view that sees discrimination as a natural and inevitable human activity that must be strictly regulated to be controlled, others believe in a market-centered view that sees group-based discrimination as a taste or rational reaction to others' tastes that can be eliminated or controlled through market means. We also discuss how ordinary empirical tests of these competing positions should be replaced with adversarial empirical collaborations. *See* Philip E. Tetlock & Gregory Mitchell, *Unconscious Prejudice and Accountability Systems: What Must Organizations Do to Prevent Discrimination?*, *in* 29 RESEARCH IN ORGANIZATIONAL BEHAVIOR (forthcoming 2009) (manuscript at 9-10, 42, on file with the Hofstra Law Review).

72. *See, e.g.*, Samuel R. Bagenstos, *The Future of Disability Law*, 114 YALE L.J. 1, 54-82 (2004) (arguing that solving the problems of disability law requires a social welfare approach). *See generally* Samuel R. Bagenstos, *The Americans with Disabilities Act as Risk Regulation*, 101 COLUM. L. REV. 1479 (2001) (discussing risk regulation in the context of disability law and arguing that public health agencies are in the best position to determine the interests of the disabled).

embrace that work.[73] But if that work turns out to be scientifically flawed, then his reformist agenda risks being brought down with it. Accordingly, it becomes important to deflect our scientific attacks and to cabin us as motivated defenders of the old, "perpetrator-focused" regime so that the larger project of rethinking antidiscrimination law can continue unscathed.[74]

In trying to rescue his value-driven agenda from its association with a youthful science that may turn out to look very different as it ages, Bagenstos ignores the underlying facts and reads into our article normative claims that do not exist.[75] Acceptance of our scientific criticisms does not depend on the acceptance of any particular normative assumptions about antidiscrimination law. Some may find our criticisms supportive of an intentional-level model of discrimination regulation and unsupportive of a move to a public-health model of discrimination regulation, as Bagenstos fears, but that would be a consequence of our scientific critique, not a cause of it.

VI. CONCLUSION

Professor Bagenstos contends that we believe that implicit bias is less morally or legally objectionable than explicit bias. But our argument was never a normative one—with the noteworthy exception that we do think it wrong to answer public policy questions with empirical research that addresses other questions. Our central argument was simply that implicit bias research fails to answer too many key questions and remains too open to external-validity challenges to shoulder the heavy descriptive and prescriptive burdens being placed on it.[76] Legal scholars invoking implicit bias research to reform the law may well have noble

---

73. *See, e.g.*, Bagenstos, *supra* note 11, at 5-10 (discussing implicit bias research and the need for a structural approach to combat this bias); Samuel R. Bagenstos, *Trapped in the Feedback Loop: A Response to Professor Days*, 49 ST. LOUIS U. L.J. 1007, 1009 (2005) ("Intentional discrimination clearly remains an important problem, but there is an emerging consensus that implicit or unconscious bias is becoming a more significant contributor to continuing workplace inequalities.").

74. *See* Bagenstos, *supra* note 2, at 488 (equating our views with "what Alan Freedman [sic] called the 'perpetrator perspective'"); *id.* at 491 ("Mitchell and Tetlock's argument likely *will* resonate with many lawyers, judges, and policymakers. The justification that best fits antidiscrimination doctrine is one that focuses, not on the individual fault of the discriminator, but on the social harms of discrimination. But most people—even most policy actors—have not internalized that justification.") (citation omitted).

75. We do not address here every detail of our arguments that we believe Bagenstos gets wrong in his effort to negate our arguments. We would simply ask the reader to read each of our original arguments carefully and in light of any clarifications we have offered here.

76. Thus, we fully agree with Bagenstos that the implicit bias research program should proceed, but applications of this work should proceed with caution. *See* Bagenstos, *supra* note 2, at 491.

intentions, but the success of law guided by social scientific research depends on the validity of the underlying research and its proper application, not on the nobility of the reformer's intentions.

For those legal scholars whose normative positions on antidiscrimination law do not hinge on the scientific validity of the implicit bias research program—a group that we believe now includes Bagenstos—our scientific concerns about implicit bias research should not matter. Yet Bagenstos expends much effort trying to reduce our scientific concerns to normative posturing, in hopes, he candidly admits, of blunting any impact our critique might have on the reform agenda he supports. We understand his concern that our critique of implicit bias research may impede reformist efforts now aligned with the implicit bias research program, but shooting the scientific messenger to protect his normative agenda is not the proper approach. Although it may be more frustrating for reformers in a hurry, the more sensible approach is to be wary when social scientists who share one's reform agenda offer research findings that dovetail conveniently with that agenda[77]—because the facts may ultimately prove less convenient than originally portrayed.

77. *See, e.g.*, Beth Potier, *Making Case for Concept of 'Implicit Prejudice': Extending the Legal Definition of Discrimination*, HARV. U. GAZETTE, Dec. 16, 2004, at 9, *available at* http://www.news.harvard.edu/gazette/2004/12.16/09-prejudice.html (describing the legal reform project of which leading IAT researchers are a part).