Title

Implicit-Bias Remedies: Treating Discriminatory Bias as a Public-Health Problem.

Journal

Psychological Science in the Public Interest, 23(1)

Authors

Greenwald, Anthony
Dasgupta, Nilanjana
Dovidio, John
et al.

Publication Date

2022-05-01

DOI

10.1177/15291006211070781

Peer reviewed

Psychological Science in the Public Interest
2022, Vol. 23(1) 7–40

DOI: 10.1177/15291006211070781
www.psychologicalscience.org/PSPI


Implicit-Bias Remedies: Treating Discriminatory Bias as a Public-Health Problem

Anthony G. Greenwald[1], Nilanjana Dasgupta[2], John F. Dovidio[3], Jerry Kang[4], Corinne A. Moss-Racusin[5], and Bethany A. Teachman[6]

[1]Department of Psychology, University of Washington; [2]Department of Psychology, University of Massachusetts; [3]Department of Psychology, Yale University; [4]School of Law, University of California, Los Angeles; [5]Department of Psychology, Skidmore College; and [6]Department of Psychology, University of Virginia

Abstract

Accumulated findings from studies in which implicit-bias measures correlate with discriminatory judgment and behavior have led many social scientists to conclude that implicit biases play a causal role in racial and other discrimination. In turn, that belief has promoted and sustained two lines of work to develop remedies: (a) individual treatment interventions expected to weaken or eradicate implicit biases and (b) group-administered training programs to overcome biases generally, including implicit biases. Our review of research on these two types of sought remedies finds that they lack established methods that durably diminish implicit biases and have not reproducibly reduced discriminatory consequences of implicit (or other) biases. That disappointing conclusion prompted our turn to strategies based on methods that have been successful in the domain of public health. *Preventive measures* are designed to disable the path from implicit biases to discriminatory outcomes. *Disparity-finding* methods aim to discover disparities that sometimes have obvious fixes, or that at least suggest where responsibility should reside for developing a fix. Disparity-finding methods have the advantage of being useful in remediation not only for implicit biases but also systemic biases. For both of these categories of bias, causes of discriminatory outcomes are understood as residing in large part outside the conscious awareness of individual actors. We conclude with recommendations to guide organizations that wish to deal with biases for which they have not yet found solutions.

Keywords

implicit bias, systemic bias, public health, disparity finding, prevention

Figure 1 plots, over a 50-year period, appearances in English-language books of six discrimination-related terms. This historical record reveals a 40-year dominance and subsequent decline of "intentional discrimination" relative to the other five. The 1970 to 2010 dominance of "intentional discrimination" likely had roots in Title VII of the Civil Rights Act of 1964, which declared it illegal to discriminate on the basis of "race, color, religion, sex, or national origin." Especially in the first several decades after Title VII became law, its phrase "on the basis of" was treated by most courts in the United States as requiring evidence of a decision maker's *intent* to discriminate. The decline in uses of "intentional discrimination" after 2000, combined with rises in the use of three other terms—"unconscious bias," "implicit bias," and "systemic racism"—signal a substantial change in both scientific and public understanding of discrimination.

Corresponding Author:
Anthony G. Greenwald, Department of Psychology, University of Washington
Email: agg@uw.edu

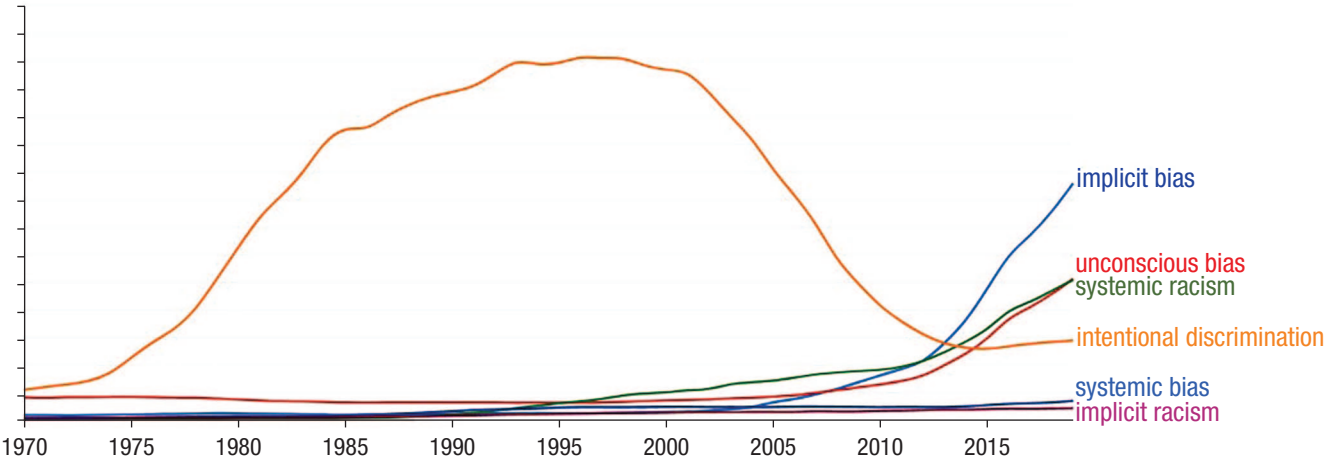

Fig. 1. Usage, from 1969 through 2019, of six concepts prominent in scholarly understanding of intergroup discrimination. This plot was produced in Google Ngram (https://books.google.com/ngrams/) by entering the six two-word terms, separated by commas, into the Ngram Viewer's search box.

Implicit bias was developed in psychology as a label for mental associations that, when triggered by demographic characteristics such as race, gender, or age, can influence judgment and behavior. The "implicit" modifier marks two characteristics of implicit bias: It is a bias that (a) manifests (and can be measured) indirectly and (b) can operate without those who perpetrate discrimination needing to be aware either of their biased associations or of the role those associations play in guiding their judgment and action.

Explicit bias (not shown in Fig. 1, but used less than "systemic bias") identifies intergroup attitudes and stereotypes of which their possessors are aware. This awareness allows for direct measurement using self-report (e.g., survey) measures. Before the first use of "implicit bias" with its current meaning (Greenwald & Banaji, 1995), "unconscious bias" was used in legal scholarship on discrimination, even if without an established scientific understanding of what "unconscious" meant in the legal context. "Unconscious bias" continues to be used in legal scholarship and elsewhere as an approximate synonym of implicit bias.

Within a few years after first publication of the Implicit Association Test (IAT; Greenwald et al., 1998), the IAT's most active developers stopped using the words "prejudice" and "racism" to describe implicit biases that could be measured with the IAT. The reason for this change was that nothing about the IAT's procedure should prompt research subjects, while their classification latencies are being measured, to bear in mind the animus (hostility or antipathy) that is a central ingredient in most definitions of racism and prejudice (see 24 of these in Greenwald & Pettigrew, 2014).[1] In contrast with IAT measures, most self-report measures of (explicit) racial attitudes oblige subjects to actively contemplate hostile or disparaging sentiments about out-groups while reporting their agreement or disagreement with statements of those sentiments. Figure 1 suggests that researchers' early choice not to use "implicit racism" may have kept use of that term to the low level it presently has.

Like "implicit racism," "systemic racism" combines a first word that implies no intent (systemic) with a second one that implies hostile intent (racism). So that this article's treatment will not be hampered by this potentially confusing juxtaposition, we proceed by using "systemic bias" where many others continue to use "systemic racism." Although "systemic bias" already has established usage (see Fig. 1), its usage is vanishingly small in comparison with that of "systemic racism." We use "systemic bias" to denote societal structures and processes that create, sustain, or exacerbate intergroup inequities without hostile intent. This includes phenomena for which the terms "institutional bias" and "structural bias" are also being used (unobjectionably). Systemic biases reside within the social system, not necessarily in the thought and decision processes of the actors who occupy society's roles that produce the discriminatory consequences denoted by "systemic bias."

As an example of systemic bias that occurs with no intent to harm, consider the home-buying decision of a family that happens to be White. If that family seeks high-quality public education—which depends on school funding, which depends on the school district's tax revenues, which often depend on local property

values, which themselves are influenced by a possibly long past history of racial segregation—that family will likely find a home in a neighborhood that is disproportionately White. Although the White family is not making their home-buying decision on the basis of race, the choice they make will likely be to invest in a predominantly White community, which will help to perpetuate existing residential racial segregation.

Preview

The concepts described and defined in this introduction will be used in the following five sections, headings of which are given here with only brief elaboration. Correctible Misunderstandings of Implicit Bias presents and corrects misunderstandings that have been propagated in public media and in some scientific treatments. What Is Known About Implicit Bias presents well-established findings of research, minimizing technical details. Research on Remedies for Implicit Bias reviews research on methods to remediate implicit biases, including the two methods that have received greatest research attention: experimental mental-debiasing interventions and group-administered trainings. We find inadequate evidence for effectiveness of either of these two approaches, a conclusion that motivates our presentation in Treating Discriminatory Bias as a Public-Health Problem, which describes the usefulness of remedies modeled on effective public-health strategies. Recommendations describes four strategies expected to reduce discriminatory consequences of implicit and systemic biases, concluding with an organizational self-test to assess the extent to which an organization has already adopted remedial approaches consistent with this article's recommendations.

Correctible Misunderstandings of Implicit Bias

Media descriptions of psychological research on implicit bias have created public awareness that discrimination can be perpetrated by persons who lack intent to discriminate. These media presentations often describe implicit bias as a recognized cause of disparities associated with differences in race, gender, ethnicity, age, disability, socioeconomic status, sexual orientation, and other demographic characteristics. Accompanying the public-education value of many of these media presentations, there have been some problematic side effects in the form of misunderstandings, which are described and corrected here. As will be seen, scientists are not free of responsibility for occasional misunderstandings.

Misunderstanding 1: The IAT and other indirect measures assess prejudice and racism

Correction: Indirect measures capture associative knowledge about groups, not hostility toward them

This misunderstanding surfaced early, appearing in a few research reports by those working most actively to develop the IAT as a research procedure. As noted in our introduction, active efforts to correct this misunderstanding were made as soon as it became apparent that it was a mistake to equate implicit bias with racism or prejudice.

A related misunderstanding is that good people do not possess implicit biases. To the contrary, the mental associations that constitute implicit biases are unavoidably acquired from the cultural atmosphere in which one is immersed daily. This cultural immersion includes literature, visual entertainment media, and audio and print news media and is also embodied in long-established practices of many public and private institutions, including the gender typing and race typing associated with many occupations. Short of imposing severe social deprivation, shielding children from their cultural environment appears not only unachievable but also possibly quite undesirable because valuable cultural content would be lost along with the stereotypes that one might avoid. There is no evidence that scrupulously nondiscriminatory beliefs and practices of parents effectively shield their children from passive acquisition of the stereotypes and attitudinal associations that pervade the larger societal environment.

The foregoing notwithstanding, a connection between implicit race bias and explicit forms of racism must be informed by the striking evidence for large-scale decline in explicit race bias in the United States and elsewhere during the second half of the 20th century.[2] Recent events (in 2020 and 2021) have made it widely apparent that hostile racism has not disappeared in the United States, regardless of what survey data reveal. Nevertheless, it is widely understood that the United States is currently characterized by a systemic form of White-favoring bias that is embodied in social institutions, policies, and practices. If one accepts the plausible and widely held assumption that implicit biases arise from widely shared, lifelong exposure to a surrounding culture (i.e., a social system), then it may be reasonable to understand implicit bias as one of the forms in which systemic bias occurs.

Misunderstanding 2: Implicit measures predict spontaneous (automatic) behavior but do not predict deliberate (controlled, rational) behavior

Correction: Implicit measures predict both spontaneous and deliberate behavior

Responsibility for this misunderstanding belongs more to scientists than to journalists. The origin of this belief was Fazio's (1990) influential two-process theory, which proposed that attitudes could operate both via a spontaneous/automatic route and via a more effortful/deliberative route. The prediction that many took from Fazio's theory (even though Fazio did not) was that implicit (presumed automatic) measures should therefore predict spontaneous but not deliberate behavior and that explicit (self-report) measures should predict deliberate but not spontaneous behavior (e.g., Asendorpf et al., 2002; Dovidio et al., 2002; Payne & Gawronski, 2010). The portion of this prediction having to do with explicit measures—that they predict deliberate, but not spontaneous, behavior—has been substantially confirmed in three meta-analyses (Cameron et al., 2012; Greenwald et al., 2009; Kurdi et al., 2019). However, the same three meta-analyses also found that the portion of the prediction having to do with implicit measures was not supported. All three meta-analyses found that implicit measures were equally effective in predicting both deliberate and spontaneous behavior.

Misunderstanding 3: Implicit and explicit biases are unrelated to each other

Correction: Implicit and explicit biases are almost invariably positively correlated

Not knowing more than that implicit race biases are more pervasive than explicit race biases, many (including many psychologists) have assumed that implicit and explicit biases are unrelated. This assumption has now been discredited in six substantial meta-analyses that examined correlations between implicit and explicit measures of a wide variety of attitudes and stereotypes. The finding in all six meta-analyses was that parallel implicit and explicit measures of individual attitudes or stereotypes are almost invariably positively intercorrelated (for a review, see Greenwald & Lai, 2020). These same meta-analyses also established that correlations of implicit intergroup attitudes and stereotypes (implicit biases) with parallel self-report measures are smaller than are correlations of implicit and explicit measures in nonintergroup domains, such as political attitudes and consumer attitudes (Greenwald et al., 2009). The finding that average implicit–explicit correlations are weaker for intergroup attitudes and stereotypes than for other attitudes and stereotypes has been widely misconstrued as indicating that implicit–explicit correlations for intergroup attitudes and stereotypes are approximately zero. In actuality, these intergroup correlations are almost invariably numerically positive. A widely suspected explanation of the smaller magnitude of intergroup (than other) implicit–explicit correlations is that explicit (self-report) measures of intergroup attitudes have additional influences, such as self-presentation concerns, that do not affect implicit measures.

Misunderstanding 4: It is scientifically established that long-established implicit biases are durably modifiable

Correction: With only occasional exceptions, experimental attempts to reduce long-established biases have not found that they are durably modifiable

The term "debiasing" is frequently used in discussions of remedies for implicit bias. In the case of implicit biases, debiasing has two goals: (a) reducing scores on measures of implicit biases (mental debiasing) and (b) reducing discriminatory behavior that might result from implicit biases (behavioral debiasing). Most experimental studies have been limited to the mental-debiasing goal, which is almost always measured in the same session as administration of the intervention being tested. As a result, there is very little evidence for durability of modifications in implicit biases produced by interventions.

Between the publication of a literature review by Blair (2002) and a collection of 17 experimental studies by Lai et al. (2014), empirical findings from studies of implicit-bias-reducing interventions were widely interpreted as supporting a conclusion that implicit biases are malleable (meaning durably modifiable). However, almost all of the mental-debiasing interventions in these numerous published reports were conducted in single-session studies, and posttests were typically obtained within 15 min after relatively brief interventions. Understanding of these results changed after Lai et al. (2016) compared effects from identical interventions as measured in within-session posttests and in posttests delayed by a day or more. The study established that

interventions found reliably to be effective when tested within the intervention session showed no effects when the posttests were delayed. This evidence and its interpretation are reviewed later under the heading Research on Remedies for Implicit Bias.

Misunderstanding 5: Group-administered procedures (often called antibias or diversity training) that are widely offered to reduce implicit race (and other) biases are effective methods of mitigating discriminatory bias

Correction: Scholarly reviews of the effectiveness of group-administered antibias or diversity-training methods have not found convincing evidence for their mental or behavioral debiasing effectiveness

Outside the laboratory, many who identify themselves as trainers offer group-administered interventions that they describe at least in part as designed to produce mental or behavioral debiasing. The absence of evidence for the effectiveness of either of these types of debiasing from group-administered trainings is understandable, given that few trainers have had professional training in the skills needed to design and conduct useful evaluations of the trainings they offer and that conducting such evaluations is rarely requested by their organizational clients.

This article's authors' encounters with descriptions of group-administered training strategies yielded the following list of advertised mental-debiasing methods: (a) exposure to counterstereotypic exemplars of members of stereotyped groups; (b) instruction to form and to remember intentions to avoid bias; (c) advice to act slowly when making decisions that might be biased (e.g., pausing to think, meditating); (d) learning about the existence and pervasiveness of implicit biases; and (e) discovering one's own implicit biases by taking available online IAT measures. With the exception of counterstereotypic exemplars and remembering intentions, these methods have not had substantial attempts at experimental confirmation of effectiveness. The two methods that have been researched have shown effects on IAT measures in within-session posttests but have not been found to yield reproducibly durable mental debiasing. Evidence for effectiveness of group-administered training is considered in detail later in this article in under the Research on Remedies for Implicit Bias heading.

What Is Known About Implicit Bias

This section divides what is known about implicit bias into (a) plausible assertions that deliberately go a bit beyond what has been established by empirical research (three "quasiconclusions"), (b) knowledge based solidly on research evidence from studies using the IAT, and (c) critiques (both resolved and unresolved) of IAT research, followed by a brief list of important questions that await research answers.

Three quasiconclusions about implicit bias

"Quasi" can be defined as "almost, but not quite." The *causation*, *pervasiveness*, and *awareness* statements in this section are all well rooted in reproducible empirical findings. At the same time, each of these statements at least mildly exceeds what can be empirically established either now or in the near future. These shortcomings do not prevent the three statements from being useful for practitioners interested in remediation of problems in which implicit biases are likely to be involved.

Causation: Implicit bias is a plausible cause of discriminatory behavior. A large body of research on how implicit-bias measures succeed or do not succeed in predicting discriminatory judgment and behavior has been summarized in three meta-analyses (Greenwald et al., 2009; Kurdi et al., 2019; Oswald et al., 2013). These meta-analyses collectively established that implicit-bias measures reliably predict discriminatory judgments and behavior. The observed correlations (*r*s) were typically small to moderate, meaning that most were between .05 and .30. Combined over the three meta-analyses, predictive-validity correlations averaged .165. Oswald et al. argued that these correlations were too small to be of practical significance, repeatedly characterizing them as "poor predictors" (pp. 171, 179, 182, 186). (That concern is evaluated under the Controversies heading later in this section.) Greenwald et al. (2015) showed that even correlations substantially smaller than the observed aggregate of .165 were "large enough to explain discriminatory impacts that are societally significant either because they can affect many people simultaneously or because they can repeatedly affect single persons" (p. 553).

"Correlation does not equal causation" is a widely known (and valid) social science aphorism. Even so, predictive-validity correlations are consistent with a causal interpretation, and this correlational evidence can be compelling about causation when there are no plausible alternative interpretations. For implicit biases, there is at least one plausible alternative interpretation—that

implicit-bias measures are themselves shaped by the same developmental experiences that produce discriminatory behavior. Alternately stated, implicit biases and discriminatory behavior have one or more shared causes. The choice between a direct-cause interpretation and a shared-cause interpretation will likely remain unresolved until methods are available to study implicit biases as they emerge in early childhood. In the present circumstance of likely continued ignorance, the causal interpretation can be regarded as useful because it is (a) simple (parsimonious), (b) intuitive (plausible), and (c) almost certain not to be empirically refuted in the foreseeable future.

Pervasiveness: Implicit bias is considerably more widespread than is generally expected. Studies using parallel IAT and self-report (explicit) measures of the same biases have found consistently, even if not invariably, that implicit measures indicate greater attitudinal or stereotype bias than do parallel self-report measures. When the two types of measures are compared in standard-deviation units, implicit measures most often show stronger biases, measured as greater differences from neutrality. It follows that implicit biases must be possessed by many who lack explicit biases. Egalitarians who are unaware of well-established research findings might expect or assume that implicit bias is no more widespread than explicit bias and that those who lack explicit race preference (such as themselves) should not expect that they possess any implicit race preferences. Unsurprisingly, a good portion of egalitarians—persons who lack explicit race bias—are distressed to obtain a race-attitude IAT score that classifies them as having moderate or strong automatic preference for racial White relative to Black.[3]

This pervasiveness conclusion is plausible to many scientists as well as to laypersons who are familiar with some of the scientific research on implicit bias. Nevertheless, the part of the conclusion expressed as "more widespread than generally expected" cannot be evaluated empirically because there are no empirical studies of expected pervasiveness of implicit biases. For those who wish to be correspondingly cautious, the pervasiveness conclusion might instead be stated as "expressions of implicit bias are considerably more widespread than are expressions of explicit bias."

Awareness: Implicit bias may produce discriminatory behavior in persons who are unaware of being biased. Statement of this awareness belief prompts a question: How does a person become aware of possessing an implicit bias? One possible route is to have an implicit bias revealed by taking an IAT (preferably more than once). A second route is to suspect that one possesses an implicit bias after having learned about the pervasiveness of implicit biases. A third route is to use knowledge of one's explicit attitude or stereotype as the basis for guessing the parallel implicit attitude or stereotype. The consistently observed positive correlations of IAT measures with parallel self-report (explicit) measures (described later in this section) can make this third strategy moderately effective, even though implicit biases are often possessed by people who lack explicit biases (see Note 3). Hahn et al. (2014) have suggested a fourth route, hypothesizing that people have introspective access to their implicit biases. Their stated conclusion poses a challenge to identify methods that can produce a convincing finding for or against their hypothesis. (This topic is returned to under this section's concluding heading, What Is Not Yet Known.)

Validity of IAT measures

Some of the material in this section may be difficult to follow for a reader without some understanding of the IAT's procedure. Most references to IAT measures in this article are to the standard form of the IAT, which has seven sets (blocks) of trials, each of which presents a stimulus (*exemplar*) belonging to one of the IAT's two target categories or to one of its two attribute categories. Four of the seven blocks (ordinally, Blocks 3, 4, 6, and 7) present *combined tasks* in which exemplars from one pair of categories appear on all odd-numbered trials and exemplars from the other pair appear on all even-numbered trials. (A full description of the standard form of the IAT appears in Appendix A. For those who have not encountered an IAT measure, it may be even more useful to go to a website where one can try out one or more IAT measures: https://implicit.harvard.edu/implicit/.)

The respondent's only instructed task in responding during an IAT measure is to press a left key or a right key to classify each exemplar into its proper category. The same two response keys are used to classify target and attribute concepts, and correct response sides for the two target categories (but not for the two attribute categories) are switched from those used initially in Blocks 3 and 4 for the second combined task in Blocks 6 and 7. The implicit measure is determined mainly by the latency difference between these two tasks, which provides the numerator of the IAT's *D* measure, computed with a scoring algorithm established in 2003 (Greenwald et al., 2003). The standard IAT procedure is described fully in Appendix A. The scoring method is described fully in Appendix B of Greenwald et al. (2021). An interpretation of how the measure succeeds follows.

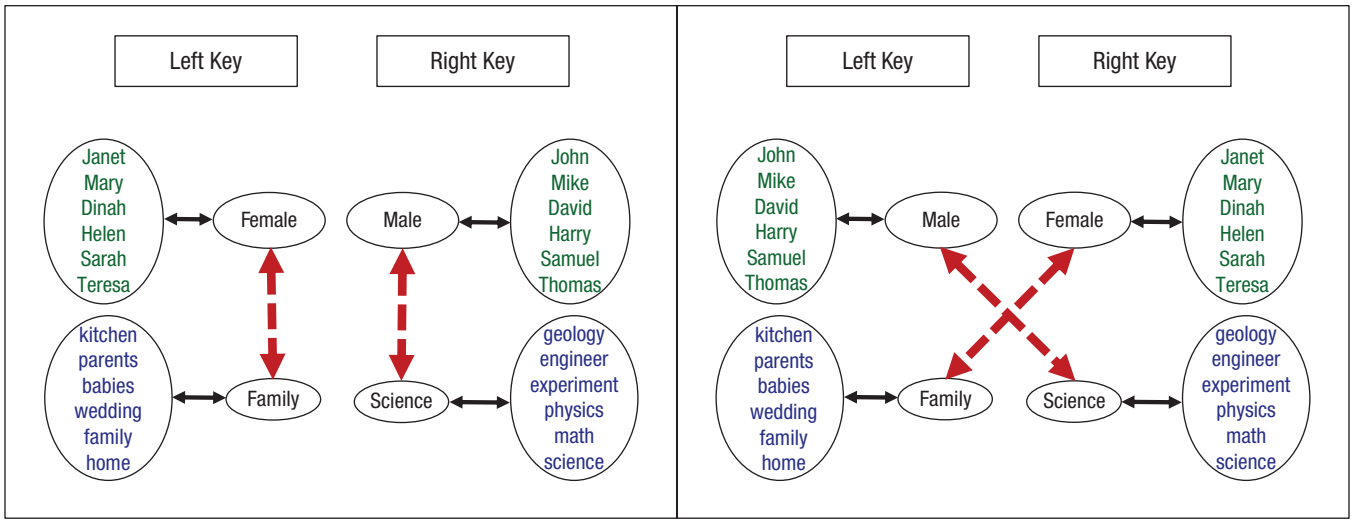

Fig. 2. Representation of associations involved in responding to an Implicit Association Test gender–science stereotype measure. The left panel shows four categories in a stereotype-consistent structure; associations link all categories and exemplars for the instructions that request a response to each key. The red arrows represent the stereotype-consistent associations. In the right panel, these associations cross between the keys, comprising a source of interference in providing the instructed responses.

A theory of what the IAT measures

In the introductory presentation of the IAT, Greenwald et al. (1998) described it as a measure of association strengths with minimal elaboration. For example:

> 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 . . . weakly associated. (Greenwald et al., 2002, p. 8; stated similarly in Greenwald et al., 1998, p. 1469)

Multiple alternatives to the association-strength interpretation offered by others are treated later in this section under the heading Critiques of Measurement and Interpretation.

Figure 2 is an initial effort (19 years after the simple statement given above) at schematic description of how association strengths are theorized to be involved in responding to an IAT procedure. Figure 2 uses the category labels and exemplar stimuli of a gender–science stereotype IAT for a person who is assumed to have associations of *male* with *science* and of *female* with *family*. The figure depicts two levels of associations: (a) associations between categories (heavier double-ended arrows) and (b) associations of exemplar stimuli to categories (lighter double-ended arrows). The associations of category labels to exemplar stimuli are assumed to have been established by many experiences of contiguity in text and speech between the exemplars and their associated category labels. The stereotype-consistent associations between categories in Figure 2 (*female* with *family; male* with *science*) are assumed to have been formed by many life experiences, perhaps including more frequent encounters with male people in scientific roles and with female people in family roles. These category-level associations are what make it easy to give the same key-press response to exemplars of both of two associated categories. When instructions assign the same key to *female* and *science*, the (often strong) female = family association will interfere with producing the instructed key press required for the *science* category on trials that present *female* exemplars. In combination, these facilitation and interference effects of association strengths produce the difference in combined-task performance speeds that is captured in the IAT's scoring method.

Construct validity of IAT measures

Cronbach and Meehl (1955) described construct validity of psychological traits as resting on a *nomological network*, which they defined as "the interlocking system of laws which constitute a theory" (p. 290). They further wrote, "Construct validation is possible only when some of the statements in the network lead to predicted relations among observables" (p. 300). The evidence for construct validity of IAT measures therefore rests on studies of correlations of IAT-measured constructs with measures of other constructs with which they should,

in theory, be correlated. This subsection summarizes evidence concerning the correlational nomological network in which IAT measures reside.[4]

IAT measures predict discrimination in judgment and behavior. After the IAT's initial publication, most IAT researchers, including the measure's creators, were withholding judgment on whether the IAT provided useful measures of implicit biases. Greatest interest was in answering the predictive-validity question: Does the IAT predict intergroup discrimination? Publications appearing in the early 2000s provided a mixture of "yes" answers (statistically significant correlations) and "no" answers (nonsignificant correlations). It was only when enough research had accumulated to conduct a quantitative (meta-analytic) review that this question began to be answered. The first meta-analysis was published in 2009 (Greenwald et al., 2009). Three of nine areas of research reviewed by Greenwald et al. involved intergroup discriminatory behavior. Subsequently, two other predictive-validity meta-analyses focused on intergroup discriminatory behavior were published (Kurdi et al., 2019; Oswald et al., 2013). The findings of all three meta-analyses revealed consistently small to moderate effect sizes of predictive-validity correlations. Greenwald et al. (2009) reported an average correlation (r) for intergroup domains of .21 (62 samples). Oswald et al. (2013) found an average predictive-validity correlation (r) of .14 (86 samples). Kurdi et al. (2019) did not report an average predictive-validity correlation for their meta-analysis. In a personal communication, Kurdi reported that this average correlation was .10 (253 samples).

Many expected known-groups differences are confirmed. *Known-groups* IAT studies are tests of differences in an IAT measure between two groups of subjects who differ in a characteristic that should be associated with either an attitude (association with valence) or an identity (association with self). Even though typically reported as a test of the difference between means for two groups, the statistical outcome is equivalent to that for a correlation in which one of the two correlated variables is a dichotomy. Expected group differences in both implicit attitudes and implicit identities are readily observed on IAT-measured attitudes when the two groups differ in (among many other things) race, gender, ethnicity, political attitudes, consumer brand preferences, religion, nationality, and university affiliations. When exceptions occur, they are informative about the consequences of stigmatized identities. The best-known case of this is the surprising absence of difference as a function of age of respondent in the strong association of age with negative valence as measured by the IAT (see top row of Table 5 in Nosek et al., 2007).

IAT attitude measures and parallel self-report measures are nearly uniformly positively correlated. This well-established conclusion is supported by findings from six large studies that reported correlations between parallel implicit (IAT) and self-report (explicit) measures for many attitudes. In a meta-analysis of 126 studies, Hofmann et al. (2005) reported that the average implicit–explicit correlation (r) was .24. In Nosek et al.'s (2007) study reporting on Project Implicit's Internet-obtained volunteer data, implicit–explicit correlations were positive for 17 IAT measures (a few of them of stereotypes), and the weighted average r was .27. Greenwald et al.'s (2009) meta-analysis, based on 155 independent samples, found average implicit–explicit correlations of .21. In 57 experimental studies, Nosek (2005) reported an average implicit–explicit correlation of .36. For 95 experimental studies conducted via Internet, Nosek and Hansen (2008) reported an average implicit–explicit correlation of .36. Kurdi et al. (2019, personal communication) found an average implicit–explicit correlation of .12 for 160 studies limited to the domain of intergroup behavior.

Because of the generally positive correlations between IAT and self-report measures, these two types of measures very often agree in direction, meaning that both have means on the same side of their zero points. Nevertheless, it is possible (but only relatively infrequently true) that means for parallel IAT and self-report measures disagree in direction. The best-known example is for measures of White versus Black attitudinal preference, for which IAT and self-report are positively correlated (e.g., $r = .207$ in Nosek et al., 2007, Table 2). This positive correlation notwithstanding, demographically diverse research samples often show explicit preferences in the pro-Black direction, on average, accompanying moderate to strong White preferences on IAT measures. Samples limited to African American respondents often show the reverse pattern of small White preferences on the race-attitude IAT, on average, accompanying strong Black preference on explicit attitude measures.

IAT findings give birth to balanced identity theory. The earliest IAT studies focused on measures of attitudes, of self-esteem, and of stereotypes, often with interest in comparing the results of these IAT measures and parallel self-report measures. An early IAT study of gender stereotypes (Rudman et al., 2001)[5] produced unexpected findings with remarkable theoretical implications. The unexpected findings were that (a) for a gender–warmth stereotype IAT, only women associated female (more than male) with warm (relative to cold), whereas men reversed the expected stereotype, associating male more than female with warm, and (b) for a gender–potency stereotype IAT, only men showed the expected stereotype of associating male (more than female) with

strong (relative to weak), whereas women associated male and female equally with strong. Rudman et al. concluded that "people possess implicit gender stereotypes in self-favorable form because of the tendency to associate self with desirable traits" (p. 1164). Because their IATs used mostly positively valenced words for "warm" and "strong" and mostly negatively valenced words for "cold" and "weak," these IATs confounded valence with trait (a practice now understood as one to be avoided; see Greenwald et al., 2021). The finding that the confound worked in different directions for men and women provided inspiration for balanced identity theory (BIT; Greenwald et al., 2002), which used a *balance–congruity principle* to predict relations among interrelated trios of measures of identity, self-esteem (or self-concept), and attitude (or stereotype). A meta-analysis recently reported by Cvencek et al. (2021) found that the novel correlational predictions of BIT's balance–congruity principle were consistently confirmed in 36 studies, involving 12,733 subjects, that had tested these predictions with both IAT and self-report measures.

Balanced identity theory predicts the frequently observed positive correlation between self-esteem and social identity. Greenwald et al. (2002) described BIT as an intellectual descendant of several affective–cognitive consistency theories that flourished in the 1950s and 1960s. BIT's most significant impact may be that it provided a new understanding of the interrelations among the central constructs of personality psychology (identities, self-concepts, and self-esteem) and the central constructs of social psychology (attitudes and stereotypes).

This new understanding was provided by BIT's balance–congruity principle. The balance–congruity principle translates to the proposition that when two categories (A and B) are associated with the same third category (C), an association between A and B should strengthen as a multiplicative function of the strengths of the A–C and B–C associations. The balance–congruity principle's name has roots in Heider's (1958) balance theory and Osgood and Tannenbaum's (1955) congruity theory. The role of self-esteem (association of self with positive valence) is important in applications of the principle because most people have strongly positive self-esteem. Consider a young girl with strongly positive self-esteem. For this girl, both female (A) and positive valence (B) are associated with self (C). It follows from the balance–congruity principle that the association between female (A) and positive valence should strengthen, producing a positive attitude toward female. Any identity (e.g., male, young, Catholic, Swedish) can replace female in the category A role, generating the expectation that the people with high self-esteem should have positive attitudes toward all of their identities. Further, strength of the positive attitude should be predicted by the strength of both the identity association and the positive self-valence (self-esteem) association. Strength of positive attitudes toward one's identities is therefore moderated by the magnitude of one's self-esteem.

The correlational findings summarized in the foregoing paragraphs constitute the "nomological network" of "predicted relations among observables" on which construct validity of IAT measures rests (Cronbach & Meehl, 1955, p. 300). Two types of correlational evidence have been described. The first is evidence establishing expected positive correlations of IAT measures with (a) measures of discriminatory behavior and judgment; (b) membership in groups that are expected to differ in attitudes, identities, or both; and (c) parallel self-report measures. This evidence supports treating IAT measures as valid measures of attitudes and stereotypes but does not bear on theoretical interpretation of IAT measures as measures of association strengths. The second type of evidence comes from tests of BIT's predictions for correlations within specified triads of measures of attitudes, identities, stereotypes, self-concepts, and self-esteem. Because those predictions depend on BIT's definition of those social–cognitive constructs as having the form of mental associations, their confirmations support that associative theoretical understanding.

Critiques of IAT measurement and interpretation

This section presents questions that have been raised about interpretations of IAT data; most of these bear on construct validity. Each section starts with brief statement of a "standard view" on a question, which is often a position contained in one or more of the early publications of IAT findings. This is followed by a paragraph briefly describing published critiques of the standard view, in turn followed by evaluation of the merits of the contrasted views.

What psychological construct is measured by the IAT?

Standard view. The IAT measures *strengths of associations* between two dimensions, each identified by a contrasted pair of categories. For the Black–White race-attitude IAT, this would be (a) a valence dimension defined by the contrast of pleasant versus unpleasant valence and (b) a race dimension defined by the contrast of racial Black with racial White.

Critique. Alternative interpretations started to appear soon after the IAT's initial 1998 publication. These included (in chronological order) *differential familiarity* of items in different categories (Ottaway et al., 2001), *criterion*

shift (Brendl et al., 2001), *figure-ground asymmetry* (Rothermund & Wentura, 2001), *task switching* (Mierke & Klauer, 2001), *salience asymmetry* (Rothermund & Wentura, 2004), *quadruple-process model* (Conrey et al., 2005), *category recoding* (Rothermund et al., 2009), *multinomial model* (Meissner & Rothermund, 2013), and *executive function* (Ito et al., 2015; Klauer et al., 2010). Further description of each is available in Greenwald et al. (2020).

Evaluation. Most of the alternative conceptions predict the same findings as the association-strength interpretation. For the few that provide bases for competing predictions (differential familiarity, salience asymmetry, executive function), empirical tests of the competing predictions have not been supportive of the critiques (see Greenwald et al., 2020). For those alternatives that make at least a subset of the same predictions as the association-strength interpretation, a generalized principle shared with the association-strength interpretation is that IAT measures capture two opposed processes or representations in a way that provides a relative strength measure that correlates with other indicators of attitudes, stereotypes, identities, and self-esteem. The exception is that the association-strength interpretation is the only theoretical interpretation that can produce the novel predictions of BIT (Greenwald et al., 2002) that were described under the preceding Construct Validity heading. Consistent experimental confirmation of the BIT predictions (reviewed by Cvencek et al., 2021) favors the association-strength interpretation.

Do IAT measures have adequate test–retest reliability?

Standard view. Average test–retest reliability (r) of .50 for IAT measures was recently reported in a meta-analysis by Greenwald and Lai (2020). Similar results had been reported in previous reviews, including Nosek et al. (2007; $r = .56$) and Gawronski et al. (2017; $r = .41$). See also Payne et al. (2017, pp. 233–234).

Critique. The IAT's test–retest reliability is too low for it to have individually diagnostic value. Payne et al. (2017) observed that "the temporal stability of [implicit] biases is so low that the same person tested 1 month apart is unlikely to show similar levels of bias" (p. 233).

Evaluation. Payne et al. (2017) did not provide data or statistical analysis to support their assertion that two administrations separated in time are "unlikely to show similar levels of bias." Data from large-sample studies of IAT bias measures are useful in evaluating their assertion. Those studies have shown that intergroup-bias measures have means averaging about 0.7 *SD* from the zero values that indicate absence of bias (data from many respondents were first presented by Nosek et al., 2007, Table 2). Using the $r = .50$ estimate for test–retest reliability of IAT measures, application of well-known statistical properties of normal distributions produces conclusions that an individual-subject observation that is 0.7 *SD* from zero at Time 1 has (a) a 76% chance of being on the same side of zero at Time 2 and (b) a 58% chance of being greater than 0.5 *SD* from zero in the same direction at Time 2.

Does the IAT assess individual differences among persons or situation differences?

Standard view. In the article that introduced the IAT, Greenwald et al. (1998) described the IAT as a measure of "individual differences in implicit cognition." This remains the standard view.

Critique. Payne et al. (2017) wrote that "most of the systematic variance in implicit biases appears to operate at the level of situations" and that "measures of implicit bias . . . are meaningful, valid, and reliable measures of situations rather than persons" (p. 236).

Evaluation. In the journal issue containing Payne et al.'s (2017) article, several of the invited commenters distributed influence on IAT measures more equally between person and situation than did Payne et al. In a later commentary, Connor and Evers (2020) pointed out that Payne et al. had interpreted the difference between correlations involving individual IAT data and large-N aggregates as indicating more "systematic variance" (p. 1331) in the group aggregates than in individual respondent data. This was incorrect, because aggregating multiple IAT scores into a mean score for the aggregate necessarily loses the (approximate) 50% of the systematic variability of individual scores that is due to between-person differences. That 50% figure, as Connor and Evers (2020) pointed out, is based on expected test–retest reliability ($r = .50$) of individual IAT scores. Statistical tests using large-N aggregates have great statistical power to detect small effects in data from which individual-person variance has been removed by aggregation. As an alert undergraduate statistics student will know, the estimated variance of aggregates of N observations can be computed by dividing the variance of individual observations by $N - 1$.

Even setting aside the article by Payne et al. (2017) and its critiques, there is much support for the individual-differences interpretation of IAT measures from well-established correlational research reviewed under this article's Construct Validity heading, including many published reports of (a) known-groups differences in IAT-measured attitudes, (b) known-groups differences in IAT measures of self-concepts or identities, (c) demonstrations of the IAT's predictive validity, and (d) positive correlations of self-report measures of attitudes with parallel IAT attitude measures. All of these findings

are consistent with the interpretation of IAT measures as measures of individual differences among persons.

Does the zero value of an IAT implicit-bias measure indicate absence of bias?

Standard view. The zero value for IAT attitude measures was described in the initial IAT publication as indicating "absence of preference" (Greenwald et al., 1998, p. 1476). As the IAT was extended to measurement of stereotypes, self-concepts, and identities, an interpretation of absence of difference between contrasted associations involved in the IAT's combined tasks was maintained. For example, the zero value of a widely used IAT measure of gender–career stereotype is assumed to indicate equal association of male and female with career (relative to family).

Critique. The absence-of-difference interpretation of the IAT's zero point has been criticized as being "arbitrary" (Blanton & Jaccard, 2006) and as "'right biased,' such that individuals who are behaviorally neutral tend to have positive IAT scores" (Blanton et al., 2015, p. 1468).

Evaluation. This critique is important because if the IAT's zero point is misplaced, a corrective relocation could substantially alter the estimated proportion of a subject population that scores in a range indicative of nontrivial bias. The critiques by Blanton and colleagues stop short of identifying an alternative location for the indifference point of IAT measures. Until recently, there was no theoretically derived basis for empirically confirming the absence-of-difference (in association strengths) interpretation of the IAT's zero value. The standard view was based on the intuition that equal speeds of performance in the IAT's two combined tasks should indicate equal strengths of the associations assumed to be drawn on in the two combined tasks (see Fig. 2). The recent development is the publication by Cvencek et al. (2021) of a meta-analysis of studies of BIT that confirms the validity of the IAT measure's zero value as indicating absence of preference. Development of the theoretical basis for these confirmations is given in both Greenwald et al. (2020, pp. 33–37) and Cvencek et al. (2021, pp. 191–194).

Are the IAT's correlations with measures of discriminatory judgment or behavior too small to be of any practical use?

Standard view. There is no standard view that identifies a correlational effect size for predictive validity that should be considered of practical importance. The criterion for significance in early IAT predictive-validity studies did not go beyond traditional use of Type I error probabilities (p values) to evaluate statistical significance.

Critique. In interpreting their meta-analysis's finding of an aggregate correlation (r) of .140 for predictive validity of IAT measures of implicit bias, Oswald et al. (2013) concluded that "the IAT provides little insight into who will discriminate against whom" (p. 188). Oswald et al. (2015) similarly concluded that "IAT scores are not good predictors of ethnic or racial discrimination, and explain, at most, small fractions of the variance in discriminatory behavior in controlled laboratory settings" (p. 562).

Evaluation. In response to Oswald et al.'s (2013) initial critique, Greenwald et al. (2015) presented statistical simulations establishing that effect sizes even substantially smaller than the aggregate magnitudes found in the three published meta-analyses of IAT predictive validity (Greenwald et al., 2009; Kurdi et al., 2019; Oswald et al., 2013) "were large enough to explain discriminatory impacts that are societally significant either because they can affect many people simultaneously or because they can repeatedly affect single persons" (p. 553). Although not contesting the validity of Greenwald et al.'s simulations, Oswald et al. (2015) restated their prior observation that aggregate predictive-validity correlations were not "large enough" to have "substantial societal significance" (p. 565).

The last three of the five critiques of IAT measures evaluated in this section appear to be largely resolved. However, the first two have potential to produce informative new empirical developments. For the first one (Are IAT measures tapping associative knowledge?), the viable alternatives have not yet generated enough empirical evidence to demonstrate that they may be superior to the association-strength interpretation of IAT measures. For the second (Do IAT measures have adequate reliability?), the current answer is "yes, but." The "but" is that research could be more efficient if reliability were stronger. A straightforward way to strengthen reliability is to use two or more administrations of an IAT measure (within a research session), at a cost of increased data-collection time. Greater research efficiency may be achievable by finding methods to reduce the inherent statistical noise of latency-based measures. Success in this endeavor would be welcomed by a sizable collection of cognition and social-cognition researchers who rely on latency measures.

What is not yet known about implicit bias

Although the first three questions listed after this paragraph have been discussed in multiple publications, they have not received confident answers. The last two have not yet received substantial attention. Investigations of all of these questions have potential both to afford new theoretical insights and to provide bases for useful applications. Some words of partial elaboration are added only for the last two. Further discussion of all of these questions is available in the consideration

of a larger set of unanswered questions by Greenwald et al. (2020, pp. 39–45):

1. How do the association strengths measured by the IAT influence social behavior?
2. When, in child development, are implicit biases formed, and what are the experiences that form them?
3. Are implicit biases introspectively (consciously) accessible?
4. What are the effects of possessing implicit stereotypes tied to one's own identities?
5. What is the effect of a person having implicit and explicit attitudes that differ?

Question 4 is suggested by an understanding of stereotype threat (Steele, 1997) as anxiety experienced by members of a stigmatized group who seek to excel in a domain in which their group is negatively stereotyped (e.g., pressure on women to perform well on a computer science test). To date, this form of conflict between an identity and a stereotype has been investigated almost exclusively in terms of explicit identities and stereotypes.

For Question 5, because of the typically positive correlation between implicit and explicit measures, large implicit–explicit discrepancies are not expected. However, there are some circumstances in which these discrepancies appear more than occasionally. The best-known example is that a substantial minority of Black respondents to the Black–White race-attitude IAT display the combination of White preference on the race-attitude IAT and strong Black preference on the parallel self-report attitude measure. There have been multiple studies with as-yet-inconclusive results seeking correlates of discrepancies between implicit and explicit self-esteem (see Cvencek et al., 2020, pp. 2–5). There has not yet been substantial research focused on consequences of other implicit–explicit discrepancies.

Research on Remedies for Implicit Bias

This section reviews methods considered as plausible or possible means of achieving mental or behavioral debiasing. The five subsections evaluate evidence on methods ranging from laboratory experimental interventions provided to one person at a time to educational approaches provided to assembled groups of participants.

Effects of experimental interventions

A variety of procedures designed to reduce implicit biases have been tested in experimental studies. The biases tested in these studies are often ones believed to be acquired early in life (for an overview of evidence for early acquisition, see Hailey & Olson, 2013). Such long-established implicit attitudes and stereotypes are mostly those associated with easily detected demographic characteristics (e.g., age, gender, and race). With the aid of IAT procedures adapted for use with preschool children, these implicit biases have been observed at ages as young as 4½ years (Cvencek et al., 2011). Unfortunately, however, the IAT method is not yet available for use earlier in childhood, when implicit biases are likely being established. Early established associations are likely not only to be strong but also to be sustained by interconnections within an associative-knowledge structure centered on the self (Greenwald et al., 2002). There is no theoretical basis for expecting these long-established implicit attitudes and stereotypes to be easily modifiable. By contrast, novel implicit attitudes or stereotypes (e.g., ones associated with previously unfamiliar persons or previously unknown groups) can easily be created and modified in experimental studies (e.g., Gregg et al., 2006). Such newly formed associations can even be reversed in valence when sufficient new information is provided (Mann et al., 2020; see also Cone et al., 2017).

Single-session experimental interventions. The first review of single-session implicit-bias interventions (Blair, 2002) included 24 studies that used the two most widely investigated measures of implicit biases, the IAT and sequential priming (Fazio et al., 1986). Only one of the 24 studies reviewed by Blair used a measure of treatment impact obtained at a time other than the intervention session, and it assessed impact just one day later (Dasgupta & Greenwald, 2001). Somewhat surprisingly, this heavy emphasis on testing the immediate impact of interventions persisted. In a 2019 meta-analysis of interventions designed to alter implicit measures, Forscher et al. (2019) found that only 38 of 598 studies (6.4%) included a posttest that was obtained other than within the session in which the intervention was administered (p. 530). Forscher et al. reported that this small proportion indicated "a lack of research interest in change beyond the confines of a single experimental session" (p. 542).

The first direct comparison of impact for immediate and delayed tests of single-session interventions was Lai et al.'s (2016) report of multiple tests involving eight interventions. Those eight interventions were ones that, in a previous collection of multiple single-session experiments by Lai et al. (2014), were found to reduce implicit race bias on immediate posttests. The eight effective interventions were in three categories: exposure to counterstereotypical exemplars, intentional strategies to overcome bias, and evaluative conditioning. Three categories

that did not produce immediate impact were taking others' perspectives, appeals to egalitarian values, and inducing emotion. Although Lai et al.'s 2016 study confirmed the previous observation of immediate effectiveness for all eight of the previously tested interventions, none of the eight displayed significant impact on posttests conducted after delays of just 1 or 2 days. Across all eight, the effects on delayed posttests averaged near zero (see Lai et al., 2016, Figure 2). Their 2016 results led Lai et al. to conclude that the interventions were "not changing implicit preferences per se, but are instead changing nonassociative factors that are related to IAT performance" (p. 1013) and that the findings were "a testament to how [implicit biases] remain steadfast in the face of efforts to change them" (p. 1014).

If the immediate changes in implicit-bias measures produced by at least three categories of brief interventions are not actually changes in implicit-bias associations, those changes must have some plausible theoretical explanation. We are aware of one plausible explanation: The observed ephemeral changes on implicit-bias measures may be *priming* effects of types that have been well established in social psychology since the 1970s. Social-psychological priming studies typically use brief interventions that activate a subset of representations associated with a mental category—for example, female scientists (a subset of females) or Black entertainers (a subset of all Black persons). The priming hypothesis is that the activated subcategory temporarily replaces a superordinate category that would have been activated but for the effect of the intervention. The activated subcategory can then be responsible for the observed temporary changes in IAT or other indirect measures of bias.[6]

Interventions based on the contact hypothesis. Pettigrew and Tropp's (2006) meta-analysis found broad support for the hypothesis that contact between members of two groups should be associated with increased liking or decreased disliking between the two groups. Largely because all of the 515 studies of intergroup contact research reviewed by Pettigrew and Tropp were reported before 2001, their review included no studies using implicit measures of attitudes or stereotypes. Because only 5% of the 713 independent samples analyzed by the authors were true experiments (i.e., with random assignment to contact versus control conditions; Pettigrew & Tropp, 2006, p. 755), the causal role of contact was indeterminate for the great majority of studies in their review.

Fortunately, more than a dozen intergroup-contact studies using dependent measures of implicit bias have been published since Pettigrew and Tropp's review. These studies have found that more frequent or more favorable intergroup contact is associated with lower levels of implicit bias. However, most of these studies were correlational, either cross-sectionally (e.g., Turner et al., 2007) or longitudinally (e.g., Onyeador et al., 2020). To avoid the causal ambiguity of nonexperimental studies, we limit attention here to five studies that used experimental designs.

Turner and Crisp (2010) found that an intervention of imagined interaction, either with an elderly person or with a Muslim person, produced a lower level of implicit bias than was observed in their control conditions. Because their only posttest was within the same session as the intervention, there is insufficient basis for treating their observed treatment effect as a durable change in an implicit bias. This limitation did not apply to Vezzali et al.'s (2012) experiment with Italian fifth-grade students. After a 3-week intervention using imagined interactions with immigrants, their delayed posttest found that children who received that intervention displayed significantly lower implicit bias than did children in their control condition. Unfortunately, a subsequent replication study (Schuhl et al., 2019) did not reproduce Vezzali et al.'s finding of change remaining evident on a delayed posttest. Another imagined-contact study found reduced implicit bias at the end of the intervention session only for subjects who were initially high in prejudice (West et al., 2017); another study found that imagined contact reduced implicit bias in a second session 24 hr after the experimental manipulation, but only when implicit-bias feedback was part of the intervention (Pennington et al., 2016). In summary, imagined contact has shown encouraging effectiveness in attenuating implicit bias under a limited set of conditions, but the robustness of that durability has not been established.

One of the most interesting and ambitious experimental tests of the contact hypothesis on implicit-bias reduction was a longitudinal study of White college students who were randomly paired with either a White or an African American dormitory roommate at the beginning of their first year in college (Shook & Fazio, 2008). The dependent measure of implicit racial attitudes was a sequential priming measure, administered both in the first 2 weeks (pretest) and in the final 2 weeks (posttest) of the 3-month academic term. The posttest revealed a significant difference between the two roommate conditions: The African American–roommate condition showed a significantly ($p = .04$) greater reduction in implicit race bias from pretest to posttest relative to the White-roommate condition. Given the high interest value of this result, it is disappointing that this finding has remained unreplicated for more than a decade. Replication would provide valuable support for the conclusion that prolonged intergroup racial contact enables durable reduction of implicit race bias.

Multisession laboratory interventions. Effects on implicit race biases of interventions extended across time have been reported in a small number of laboratory studies. The first of these (Devine et al., 2012) tested the effectiveness of bias-reduction training administered in six sessions over a 12-week period. This study reported "the first evidence that a controlled, randomized intervention can produce enduring reductions in implicit bias" (p. 1271). Disappointingly, the finding was not reproduced in a replication study from the same laboratory that tripled the initial study's sample size (Forscher et al., 2017). The authors reported that the 2017 study's multisession intervention produced effects on self-report measures, suggesting that the intervention had produced durable changes in "knowledge of and beliefs about race-related issues" (Forscher et al., p. 133).

In a two-session experiment, Stone et al. (2020) sought to reduce first-year medical students' implicit negative stereotype of Hispanic Americans[7] as medically noncompliant. Each of two workshop sessions incorporated multiple implicit-bias-reduction strategies, along with active-learning exercises relevant to medical settings. A posttest IAT was administered 3 to 7 days after the second workshop. On the IAT pretest, White medical students and those from non-Hispanic racial and ethnic minority groups significantly associated medical noncompliance with Hispanic American (relative to White American) ethnicity. In results comparing preworkshop and postworkshop measurements, White medical students showed a significant decrease in the implicit noncompliance stereotype of Hispanic patients, whereas Hispanic medical students and those from other racial/ethnic groups did not.

Interventions in field settings

Relatively few investigations of bias-reducing interventions have been conducted in settings that allow assessment of effectiveness beyond the laboratory.

A series of field experiments conducted by Dasgupta and colleagues was unusual in embedding interventions in everyday situations and in administering those interventions in person rather than via computer. These characteristics may explain why these studies, described in the next three paragraphs, found durable effects that have otherwise been extremely difficult to obtain.

In their study investigating possibilities for remedying the underrepresentation of women in professional leadership roles, Dasgupta and Asgari (2004) took advantage of geographical proximity of two similarly sized liberal arts colleges—one an all-women's college, the other a coeducational college. Their study of women's implicit gender–leader stereotypes began at the start of the women's first year on campus. At the start there were 82 women participants, 41 at each school. One year later, 63% of the sample was available for posttests, which revealed that the male = leader stereotype had been significantly weakened for women at the women's college, an effect that was strongest for those who had greatest exposure to female faculty. At the coeducational college, where the women had greater exposure to male professors in their science, technology, engineering, and math (STEM) classes, an opposite effect was observed. The male = leader stereotype had strengthened.

A second longitudinal study by Dasgupta's group examined whether contact with women professors of mathematics would change students' implicit attitudes toward math and their implicit math self-concept (Stout et al., 2011). Women students whose calculus course was taught by a female (rather than male) professor showed significantly more positive implicit attitudes toward mathematics and stronger implicit identification with the discipline (self–math association). Men's responses were not affected by variations in professors' gender. This difference remained apparent 3 months later—it was durable.

A third longitudinal field experiment from the same laboratory (Dennehy & Dasgupta, 2017) investigated consequences of randomly assigning first-year female engineering majors to a mentor who was a male senior engineering student, a mentor who was a female senior engineering student, or no mentor. Explicit measures of belonging, self-efficacy, and threat found statistically significant benefits of having a female mentor, both during the first-year mentoring experience and 1 year after the mentoring had ended. IAT results were directionally consistent with these findings—in some cases statistically significant, in other cases marginal.

Large-scale field experiments

Carnes et al. (2015) reported a study participated in by 2,290 faculty members from 92 medicine, science, or engineering academic departments at the University of Wisconsin. Faculty members in the 46 departments that were randomly assigned to the "gender-bias habit-reducing training workshop" condition received a 2.5-hr training session administered to their department unit. (Content of the workshop is described in detail on pp. 222–223 of Carnes et al., 2015.) These faculty members received a survey that assessed 13 outcome measures 2 days before, 3 days after, and 3 months after their department's workshop. Each control department received the same surveys at the same times as a matched workshop-receiving department. The results showed no impact of the workshop on the IAT-measured gender–leader stereotype. However, significant benefits of the intervention

(treatment vs. control comparisons) appeared on self-report measures of (a) confidence in personal ability to behave in gender-equitable fashion and (b) increased awareness of personal gender bias. A follow-up report by Devine et al. (2017) compared the university's hiring rates of women in the 2 years before the workshop with those in the 2 years after the workshop. This analysis found "modest [$p = .07$] evidence that, whereas the proportion of women hired by control departments remained stable over time, the proportion of women hired by intervention departments increased" (p. 213).

Chang et al. (2019) reported results of a large field experiment at a multinational corporation. The corporation recruited 3,016 employees, in 63 countries, to complete an hour-long online diversity-training session in which "participants learned about the psychological processes that underlie stereotyping and research that shows how stereotyping can result in bias and inequity in the workplace, completed and received feedback on an Implicit Association Test assessing their associations between gender and career-oriented words, and learned about strategies to overcome stereotyping in the workplace" (p. 7779). Participants were randomly assigned to three conditions: gender-bias training, general-bias training, or a control training focused on psychological safety and active listening. Measures of training effectiveness included attitudes measured at an end-of-training survey and unobtrusively measured workplace behaviors observed for several months after the training. The workplace behaviors included nominations of fellow employees for service awards and volunteering to informally mentor junior employees over coffee.

Chang and coauthors sought to demonstrate improvements of attitude and behavior supporting gender equality among training recipients who, at baseline, no more than weakly supported advancement of women's careers. They did not find this desired impact on behavior, but they did find some increase in self-reported attitudes favorable to gender equity. The only effect on behavior (but not much effect on attitude) was found for women in the United States who initially were most supportive of women's careers (Chang et al., 2019, p. 7781). The researchers concluded that stand-alone trainings such as the one they used are not likely to be "solutions for promoting equality in the workplace, particularly given their limited efficacy among those groups [especially men] whose behaviors policymakers are most eager to influence" (p. 7778).

Paluck et al. (2021, pp. 550–553) reviewed seven other large-sample field experiments that used novel and imaginative measures (although no implicit measures) to assess behavioral consequences of experimental interventions. Paluck et al. evaluated the designs and executions of these studies very favorably but characterized their takeaways, collectively, as "sobering" because "effects are often limited in size, scope, or duration" (p.553). In particular, effect sizes were described as being "much smaller than those reported on average in the corresponding laboratory literature using theoretically similar interventions" (p. 553). On a more positive note, Paluck et al. observed that "the prejudice-reduction interventions often seem more successful at changing discriminatory behaviors than at reducing negative stereotypes or animus" (p. 553).

Effects of group-administered trainings

Widespread recognition of implicit bias as a contributor to discriminatory disparities has prompted the development of many commercially provided training programs that are offered to large organizations as ways to overcome undesired consequences of implicit (and other) biases. These offerings often have three components: (a) defining implicit bias as a source of unintended discrimination, (b) describing the pervasiveness of implicit biases, and (c) advocating remedial strategies. Educational components (a) and (b) often draw accurately on scientific understanding. However, the remedial component (c) is generally not an evidence-based procedure and is typically administered with no follow-up effort to document effectiveness.

Trainers' advocacy of unsubstantiated remedies has the potential to create an unwarranted appearance that the organization receiving the training is operating in bias-free fashion. Kaiser et al. (2013) concluded that an "illusory sense of fairness" can be created when a corporation offers a program identified as "diversity training" (p. 504). This can encourage training recipients to "legitimize the status quo by becoming less sensitive to discrimination targeted at underrepresented groups and reacting more harshly toward underrepresented group members who claim discrimination" (p. 504).

The most efficient way to consider what is known about group-administered trainings is to draw on the conclusions of a few excellent scholarly reviews that have considered the impact of group-administered training either on behaviors that can affect equal opportunity among an organization's personnel or on behavior that can improve equal opportunity in hiring. We describe these important reviews, concluding each with the authors' interpretation of their findings.

Kalev et al. (2006) obtained data for 708 private U.S. companies sampled from a database maintained by the U.S. Equal Employment Opportunity Commission (EEOC). They identified three categories of diversity-oriented strategies rooted in "theories of how organizations achieve goals, how stereotyping shapes hiring and promotion, and how networks influence careers" (p. 589).

The three categories were (a) group-administered training, (b) networking with mentoring, and (c) establishing organizational responsibility. Data on the 708 companies' use of seven strategies in these three categories over the time period covered by the study (1971–2002) were obtained from company human resources (HR) managers via survey interviews. Data on employee diversity as it changed over the years covered by the study were obtained from required annual corporate reports to EEOC. In addition, the authors surveyed all of the companies, interviewing company HR managers to obtain additional data on the seven strategies over the 32-year period covered by the study.

The seven types of diversity-improving practices evaluated by Kalev et al. (2006) were (a) affirmative action plans, (b) diversity committees and taskforces, (c) diversity managers, (d) diversity training, (e) diversity evaluations for managers, (f) networking programs, and (g) mentoring programs. Kalev et al.'s judgments about effectiveness for these seven types of programs are quoted here from the Conclusion section of their article:

> Practices that target managerial bias through feedback (diversity evaluations) and education (diversity training) show virtually no effect in the aggregate. They show modest positive effects when responsibility structures are also in place and among federal contractors. But they sometimes show negative effects otherwise. (Kalev et al., 2006, p. 611)

Kalev et al. (2006) concluded, "Efforts to moderate managerial bias through diversity training and diversity evaluations are least effective at increasing the share of White women, Black women, and Black men in management" (p. 589). They found that "establishing organizational responsibility" was the only category of methods that gave indications of effectiveness in increasing demographic diversity. A more recent review by Dobbin and Kalev (2013) gave a summary statement very similar to that in Kalev et al. (2006): "[We] find that diversity training (offered to all employees or to all managers) has little aggregate effect on workforce diversity" (p. 268).

Other review articles in organizational psychology journals (e.g., Leslie, 2019; Mor Barak et al., 2016; Nishii et al., 2018) have found (similar to Kalev et al., 2006) a lack of clear findings for the effectiveness of diversity-training programs. The conclusion of Nishii et al. (2018) is representative:

> The pattern of results is filled with inconsistencies that severely limit our understanding of which diversity practices should be used, how they should be implemented, for what purpose, and to what effect. There is little theory that helps scholars and practitioners integrate disparate research results. (p. 38)

Paluck and Green (2009) reviewed 985 studies of prejudice reduction that had been reported between 2003 and 2008 (28% were unpublished). They sorted these into major categories of observational (nonexperimental, 60%), laboratory experimental (29%), and field experimental (11%). They regarded the nonexperimental studies as important because "the vast majority of real-world interventions—in schools, businesses, communities, hospitals, police stations, and media markets—have been studied with nonexperimental methods" (p. 345). Paluck and Green nevertheless concluded that the body of nearly 600 nonexperimental studies "cannot answer the question of 'what works' to reduce prejudice in these real-world settings," largely because only a third of those studies were quasi-experimental (i.e., including comparison groups but without random assignment), and of those 207 studies, only 12 afforded any basis for drawing conclusions about causal impact or the lack thereof (p. 345).

Paluck and Green (2009) were similarly ambivalent about the laboratory experimental studies captured in their review. They concluded that "those interested in creating effective prejudice-reduction programs must remain skeptical of the recommendations of laboratory experiments until they are supported by research of the same degree of rigor outside of the laboratory" (p. 351). Field experimental studies received the most favorable comment ("the field experimental literature on prejudice reduction suggests some tentative conclusions and promising avenues for reducing prejudice"; p. 356). However, their approval of the field experimental studies nevertheless remained equivocal:

> The strongest conclusion to be drawn from the field experimental literature . . . concerns the dearth of evidence for most prejudice-reduction programs. Few [of these] programs . . . have been evaluated rigorously. . . . Entire genres of prejudice-reduction interventions, including moral education, *organizational diversity training* [emphasis added], advertising, and cultural competence in the health and law enforcement professions, have never been [rigorously] tested. (p. 356)

Paluck and Green's (2009) concern about the lack of rigorous testing of diversity training was partly addressed by the large-scale studies by Carnes et al. (2015) and Chang et al. (2019) described under the

heading Large-Scale Field Experiments in the preceding section on experimental interventions. As described, those two studies showed partial indications of desired effectiveness at most and do not alter Paluck and Green's overall discouraging evaluation of diversity-training programs.

In their recent meta-analytic review of prejudice-reduction interventions, Paluck et al. (2021) included group-administered trainings along with other approaches. Paluck and Green (2009) had drawn attention to methodological weaknesses of prejudice-reduction studies reported through early 2008, including insufficient statistical power (small sample sizes), insufficient focus on behavioral outcomes, and lack of follow-up of effective laboratory interventions with experiments conducted in field settings. Paluck et al. (2021) provided a meta-analysis of a new set of 418 experiments reported in 309 articles published after those included in the 2009 review. Separate from their meta-analysis, they reported on "landmark studies that are noteworthy for sustained interventions, imaginative measurement, and transparency" (i.e., full reporting of method; p. 533). The 2021 article's conclusions indicated at least as much discouragement concerning evidence for the effectiveness of prejudice-reduction interventions as had the 2009 review, as indicated in this quote from the 2021 article's abstract:

> 76% of all studies evaluate light touch interventions, the long-term impact of which remains unclear. The modal intervention uses mentalizing as a salve for prejudice. Although these studies report optimistic conclusions, we identify troubling indications of publication bias that may exaggerate effects. Furthermore, landmark studies often find limited effects, which suggests the need for further theoretical innovation or synergies with other kinds of psychological or structural interventions. We conclude that much research effort is theoretically and empirically ill-suited to provide actionable, evidence-based recommendations for reducing prejudice. (p. 533).

The conclusion's mention of "troubling indications of publication bias" may be overly strong. The following description of Paluck et al.'s (2021) basis for this concern is paraphrased from two statements (p. 538 of their 2021 article and p. 15 of its supplemental appendix): A telltale sign of publication bias is a strong relationship between sample sizes and reported effect sizes; smaller studies must produce larger effect sizes to achieve statistical significance. Put differently, in the absence of publication bias, the authors should have obtained similar average effect estimates from small and from large studies. Their collection of studies displayed a powerful relationship of this kind. A linear regression of all effect sizes on sample sizes showed this relationship, suggesting that an extremely large study should produce no change in prejudice at all.

The observation that small studies need larger effect sizes than otherwise identical larger studies to achieve statistical significance is inarguably correct. The statement that small and large studies should produce the same observed effect sizes likewise applies only to "otherwise identical" studies. Caution is required in extending this generalization to collections of studies that are aggregated in a meta-analysis and vary substantially from being "otherwise identical." When researchers know enough about their specific procedures and measures to estimate expected effect sizes, they are wise to set the sample size to achieve a level of power appropriate for that expected effect size. Sample size may vary for other reasons, such as (a) early interruption of data collection (in the interest of economy or efficiency) when findings indicate only the unlikelihood that an already observed significant result will not be obtained with an originally planned larger sample or (b) extension of data collection in search of a statistically unambiguous finding when observed effect sizes are smaller than expected.[8] For these reasons, almost all meta-analyses will observe the negative relation between effect sizes and samples sizes (smaller sample sizes associated with larger effect sizes) that Paluck et al. (2021) found concerning. An entirely unobjectionable reason for finding this same relationship is that researchers who use estimates of statistical power to set sample sizes will use relatively larger samples sizes when they expect to find relatively smaller effect sizes. Thus, it is expected that almost all meta-analyses will show a negative correlation between sample sizes and effect sizes. Borenstein et al. (2009, p. 291) observed that the same negative correlation can also result from totally unobjectionable procedures, such as (a) use of small, but atypical, samples that are likely to show larger effects or (b) use of high-quality research procedures that should find expected effects with relatively small samples.

Interventions targeting attitudes and self-concepts related to clinical disorders

Many of the implicit associations measured in the clinical domain are associations with the self (e.g., associations with the self [vs. others] as panicked [vs. calm] to capture anxiety or with death [vs. life] to capture suicidality), whereas others target appetitive or aversive stimuli (e.g., associations with alcohol and approach vs. avoidance). Because many clinical problems emerge

at a young age and can persist into adulthood, it is plausible that such clinically relevant associations develop early in life.

In the area of anxiety disorders, several studies have tested the impact of various cognitive behavior therapies on posttreatment changes in implicit associations related to specific disorders. However, many of these studies assessed the impacts of multisession therapy immediately at the conclusion of the final session, but did not examine the durability of its impact on implicit associations over time (e.g., Gamer et al., 2008; Teachman et al., 2008; see also Renna et al., 2017).

Because those multisession studies used posttest IAT measures that directly followed the final treatment session, their data do not provide any estimate of the durability of impacts on implicit measures. Teachman and Woody (2003) stands out in using a delayed follow-up IAT posttest and in using an IAT that was not confounded with implicit self-esteem.[9] Their study tested the effect of small-group sessions of exposure therapy for individuals with high fear of spiders (arachnophobia). The IATs in this study measured the degree to which participants associated spiders with fear (afraid vs. unafraid), disgust (disgusting vs. appealing), danger (danger vs. safety), and valence (bad vs. good). Posttreatment IATs showed significantly reduced fear and disgust associated with spiders relative to the pretreatment baseline. Durability of this effect on implicit associations was maintained at a 2-month follow-up. However, this result has not yet been replicated. In addition, Huijding and de Jong (2007) found that a single session of exposure therapy did not affect responses on an implicit measure, which suggests that multiple intervention sessions are likely to be important in producing a therapeutic effect on implicit associations. A replication of Teachman and Woody's apparently durable effect would be highly desirable to afford greater confidence in the potential applicability of this finding.

We suspect that the repeated-exposure component of cognitive-behavioral therapy is likely to be central to changing implicit associations. Along these lines, a cognitive-therapy trial for depression that did not involve exposure therapy produced no change on an IAT (Adler et al., 2015), and results for interventions to shift implicit alcohol associations that do not involve exposure have been quite mixed, with some impressive findings (e.g., Wiers et al., 2011) and some null results (see Cristea et al., 2016; Lindgren et al., 2015). Teachman and Woody's (2003) clinical findings on implicit-anxiety change share similarities with Dasgupta and colleagues' findings on implicit self-conceptions in STEM (Dasgupta & Asgari, 2004; Dennehy & Dasgupta, 2017; Stout et al., 2011) and Shook and Fazio's (2008) findings on implicit racial attitudes.

Several common elements stand out across all these studies: They all involve interactive in-person exposure to people or stimuli that violate participants' preexisting expectancies (about successful scientists and engineers, about Black college students, and about the threat value of spiders, respectively), repeated over multiple occasions. These common elements may be critical features responsible for durable changes in implicit associations long after the intervention is over. We encourage future researchers to investigate and attempt to replicate the durable malleability of implicit associations in various topical domains by leveraging these three critical elements: (a) interactive in-person exposure that (b) violates people's preexisting expectancies and (c) is repeated in temporally spaced sessions.

Treating Discriminatory Bias as a Public-Health Problem

Mental-debiasing interventions and group-administered trainings have been widely treated as methods with the greatest promise for reducing unintended discrimination that might be caused by implicit or explicit biases. Our review in the section titled "Research on Remedies for Implicit Bias" found, however, that (a) mental-debiasing interventions that have been tested are not established as reproducibly effective in reducing long-established biases and (b) the value of group-administered antibias training appears limited to providing education that might motivate discrimination-reducing efforts but does not itself produce improvements in workplace diversity. Without concluding that these methods cannot be developed into successful efforts in the future, we turn, in this section, to successful public-health methods as models for strategies that have good prospects for success as remedies for discriminatory bias.

Mental debiasing and group-administered training are *curative* remedies, aimed at altering mental structures or processes believed to be responsible for discriminatory bias. In contrast, most public-health strategies are directed at persons who are not yet in a condition that calls for curative treatment. Public-health strategies are very often designed to *prevent* rather than to cure. Examples include fluoridation of water supplies to prevent tooth decay, vaccination to provide immunity that will prevent infection, seat belts to prevent serious injury from automobile accidents, and designations of public spaces as smoke free to prevent harm from passively inhaling others' tobacco smoke.

Epidemiology is the central academic discipline of public health. Wikipedia defines and characterizes this discipline compactly:

> Epidemiology is the study and analysis of the distribution (who, when, and where), patterns, and determinants of health and disease conditions in defined populations. It is a cornerstone of

> public health, and shapes policy decisions and evidence-based practice by identifying risk factors for disease and targets for preventive healthcare. ("Epidemiology," 2021)

In behavioral-science disciplines, it is often difficult to find scientists with epidemiologists' skills who similarly investigate the distribution patterns of discrimination, seeking to identify the most appropriate targets for discrimination-prevention efforts. It is noteworthy that medically trained epidemiologists have recently taken up this slack. One of their characteristic methods, which we here label *disparity finding*, is of great potential for use in the context of discriminatory bias.

Identifying and finding disparities

"You can't improve what you don't measure" is a well-worn maxim in applied science that captures the rationale for disparity finding. Disparities are inequalities. In the case of discrimination, disparities are inequalities in valued outcomes (benefits or harms) between demographic groups (race groups, gender groups, age groups, socioeconomic-status groups, etc.). Data sets containing measures that can document such demographic disparities exist in large organizations of all types, although those data often remain hidden. Disparity finding requires using these organizational data sets to identify demographic disparities experienced by those who either work for or receive services from the organizations that track these data.

Not all demographic disparities are discriminatory. To appraise whether an observed demographic disparity in outcomes is discriminatory requires answering two questions. First, is the difference in outcomes large enough to prompt concern about what its cause might be? Second, is the disparity due to discrimination? Attributing a disparity to discrimination requires evidence that the disparity is not explained by some difference between the two groups that itself is not plausibly a consequence of discrimination.[10] Applying bias-reducing efforts without first identifying the disparity as discriminatory can be well meaning, but it can also be inefficient, needlessly costly, and, worst, mistaken. In light of the clear importance of disparity finding to identifying discrimination, it is surprising how few organizations undertake disparity finding, especially large organizations that necessarily have large bodies of data that could unambiguously identify disparities that might warrant classification as discriminatory.

Consider a police department in a racially diverse city of medium or larger size. Should one assume that this city's police department is or is not discriminatory in its treatment of citizens in the course of its officers' routine patrolling activities? A publication by the Leadership Conference on Civil and Human Rights (2011) reviewed numerous reports of differences in outcomes of policing interactions associated with differences in the race, ethnicity, national origin, and religion of those with whom the police were interacting. The sources of these reports included the U.S. Department of Justice, the U.S. Department of Labor's Bureau of Justice Statistics, the U.S. Department of Homeland Security, multiple published surveys and law-review articles, and numerous newspaper investigative reports (Leadership Conference on Civil and Human Rights, 2011, pp. 9–19). The data in many cases were obtained in response to discrimination suits and investigations initiated by outside agencies. One might wonder why these disparities were not discovered by the police departments themselves. A likely explanation: Vigilance in identifying discriminatory disparities by police departments is not generally required by federal, state, or municipal governments. (For an important exception, see Box 1.) In this nonregulatory climate, neither police departments nor their superiors in local government are likely to be motivated to examine or report on data that are likely, on close examination, to reveal demographic disparities that can expose discriminatory disparities.

For their own administrative purposes, most large organizations maintain databases that record important outcomes to their personnel (pay, promotion, performance evaluations, and discipline) while also including data on employee qualifications (e.g., education, prior employment, and scores on aptitude or knowledge tests), along with standard demographics (usually including at least gender, race, age, ethnicity, and disability status). When a possibly discriminatory disparity is identified—for example, a male–female difference in pay or promotion—the needed next step is to determine whether that disparity has a plausible nondiscriminatory explanation. For a gender disparity in pay, a nondiscriminatory explanation can be supported if there are gender differences in objective indicators of skills, qualifications, and performance that correlate unambiguously with observed pay differences (and these differences are not themselves due to discrimination). Repeated periodically, analyses searching for demographic disparities can reveal if or when remedial efforts succeed in reducing disparities that lack nondiscriminatory explanations.

Consider also this example from health care: Marin et al. (2021) examined the use of diagnostic imaging (ultrasound, X-ray, computed tomography, MRI) for children, searching for demographic disparities. They found that Black and Latinx children received significantly less diagnostic imaging in emergency department visits than did White children. To judge whether these

Box 1. Discriminatory-Disparity Finding Illustrated: Policing in California

The 2003 book-length Institute of Medicine (IOM) report *Unequal Treatment* documented multiple demographic disparities in health care, attributing some to implicit or unconscious bias. A study of recent data (Agency for Healthcare Research and Quality, 2020) based on annual federal reviews mandated since 2003 found that "some [health-care] disparities were getting smaller from 2000 through 2016–2018, but disparities persist and some even worsened, especially for poor and uninsured populations" (p. ES1). At the time of the 2003 IOM report's appearance, the medical community's lack of awareness of these biases was suggested by the following statement, which appeared in a medical journal's critique of the report: "It is doubtful that hidden forms of discrimination are prevalent in a profession whose professional norms are set so strongly against it" (Epstein, 2005, p. S26). Indicating a similar current lack of awareness among Americans more generally, a 2019 Pew Research Center report found that 66% of adult Americans (73% of White Americans) perceived that Black Americans receive medical care at least equal to that received by White Americans (Pew Research Center, 2019).

In three other domains—banking, retail merchandise, and voting—that same Pew Center survey found that majorities of adult Americans perceived that Black Americans received treatment as good or better than that received by White Americans. (Only in criminal justice and policing were Black Americans perceived as being treated worse.) Most large organizations in the United States are sitting on large bodies of data that could be used to discover discriminatory disparities, but they generally do not report publicly on what those data reveal. The frequency with which discriminatory disparities are unexpectedly discovered the first time they are searched for indicates that it is inappropriate to assume that unexamined data harbor no evidence of discriminatory disparities.

It is, in fact, rare to find any large organization that takes steps to discover (let alone report publicly on) discriminatory disparities in its own operations. Here we describe such a noteworthy effort that is annually examining evidence for disparities in traffic stops in the largest state in the United States, including both analyses that can reveal demographic disparities and additional analyses that can test whether these disparities are discriminatory.

In 2015 the California legislature established a 19-member Racial and Identity Profiling Advisory Board (RIPA) in California's Department of Justice. RIPA was asked to report annually on demographic disparities in outcomes of stops by state and municipal police. RIPA's most recent report (Racial and Identity Profiling Advisory Board, 2021c) described disparities in stops conducted in 2019, the first full year for which data were available. Multiple disparities were evident (see pp. 35 and 50–51; also see Appendices A.1 and A.6 in Racial and Identity Profiling Advisory Board, 2021a). Drivers were Black for 16% of all stops, more than double the percentage of Blacks (7%) in California's population. For comparison, Whites accounted for 33% of stops, slightly less than their percentage (34%) of the population. In addition, rates of conducting searches for contraband (drugs or weapons) were considerably greater for drivers classified as Black (20.2%) or Hispanic (12.2%) than for White drivers (8.2%). Perhaps most remarkable is that the conclusion of a stop with "no action taken" was much greater for White drivers (85.4%) than for Black drivers (67.4%); rates were intermediate for Hispanic (79.7%) and multiracial (78.5%) drivers. Last, whereas the subjective category of "reasonable suspicion" for a stop was given as the basis for 11.7% of stops of White drivers, that figure was almost doubled (21.0%) for Black drivers. Such discretionary stops are recognized as a problematic category that is most likely to reflect racial/ethnic biases in the form of a lower threshold for suspicion for Black drivers (see Charbonneau & Glaser, 2021). Another substantial demographic disparity in the stops data was in reports of uses of force, which showed that the odds of an officer's using force during a stop were 2.45 times as great for Black and 2.18 times as great for Hispanic than for White drivers (Racial and Identity Profiling Advisory Board, 2021a, p. 57).

To identify an observed demographic disparity as discriminatory, evidence that might rule out demographic differences as the basis for the disparity must be considered. In the California traffic-stop data, two types of evidence were available. First was a type of evidence that has been observed in other investigations of police traffic and pedestrian stops: rates of discovery of contraband (drugs or weapons). Typically, these investigations have found higher rates of contraband for stopped White drivers than for stopped Black and Hispanic drivers. This was also true for the RIPA data (see Racial and Identity Profiling Advisory Board, 2021a, Appendix C.2.2.1, and 2021b, p. 3). The second type of evidence came from a "veil of darkness" analysis, which revealed that stops after sunset, when drivers' race should be less visibly identifiable, showed little evidence of the racial disparities that were otherwise noticeable. These two types of evidence established the observed disparities as discriminatory.

disparities were discriminatory, Marin et al. analyzed multiple variables that might allow the disparities to be judged nondiscriminatory, including availability and source of medical insurance and specific diagnosis. Controlling for these additional variables and others reduced but did not eliminate significant evidence for disparities—in other words, the disparities were judged to be discriminatory.

Comprehensive examination of health-care disparities and of the additional evidence needed to determine whether they are discriminatory is now provided annually in reports by the U.S. Department of Health and Human Services Agency for Healthcare Research and Quality (2020). Their most recent report is described further in Box 1, which also presents a nonhypothetical illustration of disparity finding in policing data that are now annually being collected, analyzed, and reported on by the Racial and Identity Profiling Advisory Board (2021b) in California's Department of Justice.

Most large organizations—including corporations, hospitals, governments, court systems, police departments, school systems, and universities—possess not only the types of personnel records just described but also data on their delivery of services to clientele such as students, job applicants, customers, automobile drivers, medical patients, criminal defendants, prisoners, loan applicants, and welfare clients. These databases collectively constitute a massive trove of evidence that can be used to reveal demographic disparities in provision of services, to evaluate possible nondiscriminatory explanations for those disparities, and to track changes in the observed disparities over time. Although these databases might effectively be used for these purposes, at present most organizations lack incentive to use their databases to evaluate the unintended discriminatory consequences of their operations.

Public-health approaches to ameliorating effects of unintended discrimination

On its website, the American Public Health Association (APHA) recently started reporting on identification of race discrimination as a public-health problem. As of mid-October 2021, APHA had recorded declarations by 217 U.S. governmental entities, each asserting that racial discrimination constitutes a "public-health crisis" in the United States (American Public Health Association, n.d.). These declarations appeared between 2019 and 2021; about 90% were issued in the18 months following the May 2020 murder of George Floyd. Many of the declarations suggested or encouraged action that might be taken by governments or nongovernmental organizations in the 37 states from which the declarations originated.

A list of action proposals for public-health remedies for discriminatory bias advocated by APHA appears in their report titled "Achieving Health Equity in the United States" (American Public Health Association, 2018). Those recommendations included urging increases in funding by the U.S. Congress for public-health actions, including not-yet-funded components of the Affordable Care Act, which was passed by the U.S. Congress in 2010. Also advocated were actions that could be taken by the U.S. Centers for Disease Control and Prevention (CDC) and major governmental agencies and nongovernmental organizations with health missions (e.g., "The CDC should continue to monitor progress on eliminating health disparities and health inequities and consider new methods to expand monitoring of health outcomes by socioeconomic status in addition to race/ethnicity").

APHA's recommendations for action are timely, but they may not go far enough to be helpful to leaders of many organizations that have diversity and inclusion goals. That is, they include little in the form of recommended effective policies and actions. An example is the August 5, 2020, executive directive from a state governor on the topic of addressing racism as a public-health crisis (Whitmer, 2020). That order has four sections, titled "Data and Analysis," "Policy and Planning," "Engagement, Communication, and Advocacy," and "Training." Each of the first three sections describes reasonable strategic goals, but without specifying methods known to be effective in achieving those goals. Specific procedures are described only in the "Training" section, which states that employees in all state departments and agencies should be required "to complete implicit bias training."

Paine et al. (2021) took note of the rising frequency of declarations of racism as a public-health crisis, analyzing three of them using "evaluative criteria aligned with health equity and antiracism practices" (para. 1). For these three programs, Paine et al. judged the extent to which they were actionable, financially responsible, and participatory and the extent to which they addressed structural determinants of equity. We provide our own suggested criteria for evaluating organizational discrimination-remedial strategies in our concluding Recommendations section.

Prevention: decision blinding and discretion elimination. Public-health remedies for discriminatory bias can be *preventive* (harm avoiding), *governmental* (e.g., health laws, regulations, and mandates), and *reparative* (damage fixing). These differ from the two curative remedies (mental debiasing and group-administered training, reviewed in the section titled Research on Remedies for

Implicit Bias) that have been the main foci of existing efforts to remediate effects of implicit bias. Little work has yet been done on governmental and reparative remedies for implicit bias. The remainder of this section focuses on what can be recommended in the way of noncurative remedies, especially ones based on prevention.

Only one aspect of the operation of implicit biases is known with enough confidence to provide a firm basis for preventive strategies. Common sense makes it obvious that the necessary trigger to activate an implicit bias is awareness of demographic characteristics of a person or group about whom one is making an evaluative judgment or decision. This understanding justifies two strategies for preventing discrimination that could result from implicit biases. *Decision blinding* involves procedures that prevent a decision maker from knowing the demographic characteristics of a person or group being evaluated, thus deactivating demographic triggers in the decision-making process. *Discretion elimination* includes procedures that restrict the basis for making a decision to nondemographic decision-relevant information.

Both decision blinding and discretion elimination are now in use, although infrequently. Their infrequent use is likely because many decision makers may (a) assume that they are capable of ignoring demographic information in their decision making; (b) believe that decision blinding or discretion elimination eliminates their ability to rely on their personal expertise, which they expect should be most useful when applied to all available information about the person (i.e., including demographic characteristics); or (c) succumb to organizational inertia, which can prevent managers and executives from considering opportunities to implement novel procedures.

Decision blinding. The most widely known example of decision blinding is the use of a curtain to separate auditioning instrumentalists from the members of symphony orchestra audition committees who are evaluating their performances. This blind-audition method was adopted by most major symphony orchestras in the United States in the 1970s and 1980s. The proportion of women instrumentalists who were hired by these orchestras increased dramatically, from about 20% before 1970 to about 40% after the introduction of blind auditions (Goldin & Rouse, 2000, Figure 3). When decision makers or performance evaluators do not know the demographic characteristics of those whom they are judging, it is logically impossible for the implicit biases of those decision makers to be triggered by those (unperceivable) demographic characteristics.

As already mentioned, decision blinding is used infrequently. Blinding should nevertheless be possible whenever there is a performance or a work product that can be evaluated without the evaluator encountering information that identifies the creator. Decision makers who are likely missing opportunities for blinded decision making include teachers grading the work of students, corporate managers evaluating their supervisees' accomplishments, and physicians evaluating residents, interns, nurses, orderlies, and other staff they supervise. Decision makers for whom blinding may be entirely impossible include judges and juries evaluating the veracity of defendants or witnesses in criminal cases or parties in civil cases, voters evaluating political candidates, doctors treating their patients, and selection committees reviewing applicants for upper management positions for which only a few well-known eligible applicants are being considered.

Discretion elimination. In many U.S. courts that try employment discrimination cases, discretion in personnel decision making is recognized as an enabler of discriminatory outcomes (see Hart, 2005; Heilman & Haynes, 2008). Use of discretion can problematically allow discriminatory assumptions or inferences based on an employee's demographics. Discretion can be eliminated by obliging a decision maker to rely exclusively on nondemographic criteria for evaluation. Note that this is not the same as eliminating all discretion from decision making. The remaining discretion is seen readily in the example of the blind audition, for which audition committee members are expected to exercise their discretion in evaluating the quality of audible performances. Likewise, a teacher grading an anonymous essay has considerable discretion in evaluating the essay's text. Elimination of demographic characteristics as a basis for discretion will, in principle, exclude any of the potentially discriminatory subjectivity that can otherwise occur when demographics can influence the judgment. In practice, as will be seen, discretion elimination can be difficult to achieve.

Structured interviews. Structured interviews are designed to eliminate potentially discriminatory discretion by having decision makers precommit to decision-appropriate and objectively specified qualification criteria before evaluating applicants competing for a position. The ideal structured interview (see McCarthy et al., 2010) has three components that should be administered as identically as possible to all applicants: (a) a set of questions selected because the answers should be informative about qualifications for the available position; (b) interviews conducted by asking these questions in the same fashion of all applicants, with responses fully recorded; and (c) a standard protocol for scoring interview responses to produce summary measures that have been established (empirically) as predictive of successful job performance. These ideal characteristics

are simultaneously the main virtues and the main difficulties of the structured-interview method.

When conducted in the just-described ideal form, structured interviews can be very effective. But they are difficult to implement in this ideal form, in part because achieving the ideal form can be remarkably labor intensive. When a large number of persons are being hired for the same type of position—a condition that helps to justify the effort needed to create a structured interview in the ideal form—the personnel time required to administer structured interviews includes the sum of times required to conduct and to score a single interview, multiplied by the number of applicants, in turn multiplied by the number of interviewers evaluating each applicant. Because of these demands of creation and execution, structured interviews are often implemented with shortcuts in one, two, or all three components of design, administration, and scoring. The undesired consequence is a compromised structured interview that can afford substantial discretion to interviewers and scorers. To the extent that discretion thus gains entry into nonideally implemented structured interviews, the resulting decisions are open to influence by implicit (or other) biases.

A separate challenge for the structured-interview method in its ideal form stems from its being a mechanical decision-making procedure. Writing about mechanical decision-making procedures, Kahneman and Klein (2009) observed, "the introduction of algorithms and other formal decision aids in organizations will often encounter opposition and unexpected problems of implementation. Few [decision makers] enjoy being replaced by mechanical devices or by mathematical algorithms" (p. 524).

Objective testing of aptitudes and skills. An existing widely used and often effective method of discretion elimination is to base hiring decisions and competitive school admissions on validated, objective tests of needed aptitudes or acquired skills. Objective testing is widely used in evaluating aspirants to many types of skilled occupations, including attorneys, nurses, electricians, plumbers, accountants, dentists, physicians, architects, attorneys, and engineers. Unlike structured interviews, such certification examinations are administered and scored without opportunity for graders' discriminatory discretion and without the number of applicants being a multiplier of administration and scoring effort. Objective certification examinations are expensive to produce, but the cost can be well justified when there is sufficient societal demand for people who are certified as having the tested knowledge and skills.

The existing uses of objective tests just described cannot be considered without mentioning two concerns that diminish enthusiasm for their use, both having to do with plausible sources of bias that can favor higher-status groups. First, many objective testing procedures, both in the professional certification and the competitive admissions contexts, include oral components or performance opportunities that are observed and evaluated by one or more experts in the domain being tested. If the demographic characteristics of the examinee are thereby revealed, any discretion in scoring afforded to the expert judges will partially defeat the purpose of the components that are objectively scored. Second, if the objective criteria were originally validated in a demographically nondiverse population, it should not be assumed that the evaluation is equally valid when used with demographic groups not included in the original validation studies. One basis for invalidity might be that the test assumes a knowledge base that may not have been acquired by ethnic or racial groups other than the groups with which the test was first validated. A second basis is the possibility of stereotype threat (Steele, 1997). Steele conceived stereotype threat as impairing performance on tests when (a) important outcomes can depend on the test results and (b) the test taker identifies with a demographic group that is negatively stereotyped on the aptitude or skill being assessed

Objective scoring of written materials. A second discretion-elimination possibility is development of objective scoring protocols for written materials that are used in evaluating applications for many professional positions. The main problem with this method is that objective scoring protocols are almost never available for evaluating written materials in student admissions or faculty hiring, which may include personal statements, curricula vitae, publications, and letters of recommendation. Even though it is possible to identify judges with expert knowledge of specific professional domains—for example, in obtaining expert reviews for journal submissions—those judges (e.g., journal editors and reviewers) typically have great discretion in providing their judgments (and editors in selecting reviewers), such that involvement of implicit biases in these judgments is likely unavoidable (see Greenwald, 2009).

Artificial intelligence. A recently developed third discrimination-elimination possibility is using artificial intelligence (AI) to replace subjective human decision making. At present, AI-assisted decision making must be treated as a method that awaits validation. The most widely used AI strategy is based on *machine learning*, a data-mining method that can be applied to large databases of past decisions of the type (e.g., hiring or competitive admissions) that is to be automated. The machine-learning

method produces algorithms that identify new applicants whose characteristics most closely match those of past successful applicants for the same type of position. Initial uses of machine learning for hiring or admissions decisions were found to produce algorithms that displayed gender or race biases. This undesired result was well understood as occurring because gender and race (and, likely, other) biases were contaminants of the records of past decision making. (See Bender et al., 2021, for a broad discussion of biases and other problems of machine-learning methods applied to natural language processing.) Without substantial additional developmental effort, machine-learning-assisted AI cannot be expected to avoid reproducing past patterns of bias.

The AI method of *word embedding* extracts multidimensional word meanings from large bodies of text. Using this method, Bolukbasi et al. (2016) found that these extracted meanings incorporated stereotype-based gender biases. Caliskan et al. (2017) went further to show that the word-embedding method captured meanings of demographic categories (male, female, young, old, etc.) that included biases closely agreeing with those revealed by IAT measures of implicit attitudes and stereotypes associated with those demographic categories. At present, therefore, machine-learning-assisted AI decision making is compromised by its inheritance of biases that have effectively been fossilized in the preserved records of past decisions. AI researchers can be commended for having identified this problem early in their uses of machine learning to construct decision algorithms. Although ongoing work seeks to expunge these biases computationally, current indications are that this is a task that has no easy solutions (Gonen & Goldberg, 2019).

Further considerations.

Usefulness of implicit-bias remedies for disparities caused by other forms of unintended discrimination. The disparity-finding, decision-blinding, and discretion-elimination strategies described here as remedies for implicit biases should prove equally applicable to discriminatory disparities produced by other sources of unintended discrimination, including those identified as systemic biases. Systemic biases and implicit biases have in common that they can inflict disparities in small and possibly individually nonapparent doses. These small doses can become substantial if they occur to many people or if they occur repeatedly to the same person (see Greenwald et al., 2015). When these small doses selectively affect a specific demographic group, their adverse impacts become readily noticeable when data are aggregated at the group level.

Pipeline problems. Consider a problem for which then-President Lyndon Johnson, in his 1965 commencement speech at Howard University, made this observation:

> You do not take a person who, for years, has been hobbled by chains and liberate him, bring him up to the starting line of a race and then say, "you are free to compete with all the others," and still justly believe that you have been completely fair. (Johnson, 1965, para. 12)

Johnson's metaphor compellingly dramatized how biased access to developmental experiences can cause deficits in qualifications (e.g., for a work or educational opportunity) that do not reflect deficits in inherent potential. These qualification deficits have been called "pipeline" problems, meaning that the flow of applicants to valued occupational roles has been biased by deficits in their developmental and educational opportunities.

Pipeline problems are evident when data analyses reveal (a) demographic disparities in outcomes correlated with (b) disparities in qualifications that are, in turn, correlated with (c) disparities in past educational or developmental opportunities. Attempts to repair pipeline problems identified in this fashion are typically problematic because of the likelihood that no present actor was responsible for the past deficits. In his speech, other than saying that he was addressing "this generation" of Americans, Johnson left unspecified the "you" to whom he addressed his conclusion about being "completely fair" (Johnson, 1965). Scientific understanding of how implicit and systemic biases could be responsible for past deficits does not help with the challenge of determining who should most appropriately be responsible for repairing the disparities that are the lasting residues of those past deficits.

As a postscript to this section on applying public-health methods to remediation of unintended discrimination, we take note of the frequent existence of governmental programs created to prevent, protect against, or repair harms to citizens. These include laws or regulations that institute penalties for noncompliance with preventive measures. Examples include fines for driving without fastening a seat belt, fines to businesses that do not enforce designated smoke-free zones, and legally mandated vaccination for children attending public schools. Box 2 takes note of the paucity of comparable programs to deal with discriminatory bias. Governmental protective, preventive, and reparative efforts in civil-rights domains are substantially weaker than those in multiple other domains.

Recommendations

With our scientist hats on, we know it is not our role to advise others how to do the work of discrimination reduction. However, as authors who suspect that a

Box 2. Governmental and Regulatory Remedies for Discrimination: A Road Ahead?

The best-known governmental remedies for discrimination are the series of civil-rights laws enacted by the U.S. Congress starting in 1957. For example, Title VII of the 1964 Civil Rights Act prohibited discrimination in employment on the basis of race, color, religion, sex, or national origin. Enforcement of Title VII's prohibitions of discrimination requires initiative by the person who seeks redress for discrimination. The first step is to submit a complaint to the Equal Employment Opportunity Commission (EEOC). The complaint triggers an agency investigation. Reviewing data that documented the travails of plaintiffs thus seeking redress for employment discrimination, Clermont and Schwab (2009) wrote, "results in the federal courts disfavor employment discrimination plaintiffs, who are now forswearing use of those courts" (p. 104).

Contrast this approach with that of other governmental and regulatory agencies that assume the burden of helping the injured party. Federal agencies serve citizens and others in dealing with a wide variety of potential or experienced harms, such as workplace safety (Occupational Safety and Health Administration [OSHA]), natural disasters (Federal Emergency Management Administration), air and water pollution (Environmental Protection Agency), transportation safety (Federal Aviation Administration and National Highway Traffic Safety Administration [NHTSA]), and many medical conditions (Health and Human Services, Food and Drug Administration, and Centers for Disease Control and Prevention). Unlike the civil-rights offices that deal with discrimination, these agencies—ones that deal with a wide variety of other harms—regularly provide remedies without the injured party having to pursue individual litigation. For example, OSHA can conduct inspections of worksites without advance notice and issue citations and fines, without leaving injured employees obliged to pursue the matter on their own. And NHTSA can find that an automobile safety defect exists and issue a recall, triggering broad rights of repair, replacement, or refund, then notifying affected consumers without their having to file individual lawsuits.

As just described, governmental and regulatory roles in proactively providing protective and preventive efforts exist in industrial, medical, and transportation domains. There are numerous state and local agencies that engage in parallel efforts, additionally undertaking safety inspections of commercial establishments and licensing of professionals who play roles in assuring the safety of goods and services. Not all of these activities are labeled public-health practices, but they do share the protective and preventive characteristics of public-health efforts.

A concluding question: What will be required for the types of discrimination recognized in U.S. civil-rights laws to be treated proactively and preventively, as public-health concerns are treated by many federal and state agencies? This article is not an appropriate place for advocacy, but it is nevertheless appropriate in this article to identify practices worthy of consideration if discriminatory bias were to be reframed as a public-health concern. In our understanding, the least expensive and most effective approach would be for governments to facilitate or to require the practice of disparity finding, much as was required by California's Department of Justice for traffic policing in California, as described in Box 1.

majority of this article's readers wish to reduce discrimination, we have no hesitancy in describing how those readers may put this article's scientific conclusions to use. So that this section will be self-contained, we start with definitions of four important terms used in the recommendations, in a few cases restating definitions given earlier in this article.

Four definitions

Equal treatment. After being aggregated within demographic groups, any measurable outcome can be contrasted between pairs of those groups. Equal treatment exists when statistical analysis indicates absence of differences in valued outcomes between groups (e.g., between Black and White or between Native American and European American). For an organization's employees, valued outcomes include hiring, starting salary, raises, bonuses, promotions, performance evaluations, discipline, layoffs, and terminations. For clients or service recipients, these could be value of service received or outcomes experienced, which will vary considerably depending on who the clients are—such as customers, prisoners, students, medical patients, or welfare recipients.

Demographic disparity. Demographic disparities are differences in valued outcomes between (i.e., unequal treatment of) two demographic groups. In the United States, demographic disparities are worthy of legal attention when the groups are two mutually exclusive protected

classes. United States civil-rights laws identify protected classes variably across federal and state jurisdictions but typically include groups defined by differences in age, race, color, religion, gender, national origin, physical or mental disability, sexual orientation, and veteran status.

Discriminatory disparity. A demographic disparity can be inferred to be discriminatory if it has no plausible nondiscriminatory explanation. For example, when there is a disparity between two demographic groups in average bail amounts set by judges, the disparity should not immediately be classified as discriminatory if the group having larger bail amounts has been arrested for more serious offenses. If the disparity remains when the analysis controls for seriousness of offense, it may be difficult to find a plausible nondiscriminatory explanation.

Disparity finding. The collection of activities of statistically discovering demographic disparities and determining whether they are discriminatory is what we identify as disparity finding.

Four discrimination-reduction strategies

Make disparity finding a standard practice in large organizations. A main conclusion of our section headed Treating Discriminatory Bias as a Public-Health Problem was that disparity finding is an effective contributor to bias reduction but is underused outside of health care. Although discovering a disparity does not eliminate the disparity, it makes the problem explicit, reveals the magnitude of the challenge, and can make clear who is best positioned to undo the disparity. As an example, consider the discriminatory disparity that was discovered by Marin et al. (2021) in uses of diagnostic imaging for children receiving emergency medical care. The medical community understands that this disparity can be fixed by developing an evidence-based protocol for decisions about choices of imaging technology for pediatric diagnosis, followed by monitoring for adherence to the protocol. In the case of the racially discriminatory disparities found in the Racial and Identity Profiling Advisory Board's (2021b) analysis of California traffic stops (see Box 1), the needed fixes are not so obvious, not least because California's State Department of Justice (which discovered the disparities) lacks authority to implement changes in the municipal police departments that produced the disparities.

Strategies based on bias prevention warrant greater use and greater development. The review in our Remedies section found that, despite receiving much attention by researchers and practitioners, attempts at achieving mental debiasing either via treatment interventions or via group-administered training are mostly ineffective in their present states of development. Because of this, needed remedies for unintended discrimination (i.e., discrimination on the basis of implicit biases or systemic biases) have some similarities to those for incurable infectious diseases. Incurable infections are managed primarily with preventive strategies. For unintended discrimination, two types of preventive strategies, decision blinding and discretion elimination, are known but receive relatively limited use. Research directed at developing additional preventive strategies has also been limited, perhaps because many researchers have focused more on the experimental testing of (so far unsuccessful) bias-reduction interventions.

Cautions regarding remedies described as "training." The greatest commercial development of strategies directed at undoing discriminatory consequences of unintended bias is in offerings marketed as "training." It is easy, but incorrect, to assume that a service labeled "antibias training" or "antiracism training" will solve problems that may have led leaders of an organization to be accused of race, gender, or other bias. There is considerable variety in training offerings, and at present, there is no licensing or certification of their effectiveness by any appropriate professional organization. There is also an absence of regulatory oversight. Accordingly, we offer three pieces of advice to those who consider investing in purchase of training services. The first is based on our previous observation that well-done antibias training can effectively provide education about implicit bias and systemic bias. It is therefore appropriate to ascertain that those offering the training are up to date on scientific understanding of implicit and systemic bias. Unless the trainer is a person with an established scientific reputation, this requires not only advance access to the trainer's presentation but also to the advice of someone with sufficient expertise to evaluate that presentation. Second, if the goal is to correct known inequities in treatment of staff or clients, we advise asking a training contractor to describe the research evidence that warrants confidence that the proffered training will achieve those goals. Third, any training that is implemented should be accompanied with follow-up observation to establish whether expected improvements materialize.

Organizational structure of diversity, equity, and inclusion efforts. Organizations may outsource management of diversity, equity, and inclusion (DEI) concerns or they may assign that responsibility to their own personnel. If outsourcing, the cautions advised above when contracting for training apply. There may be many ways to organize in-house DEI efforts effectively. We limit the advice under this heading to just one proposal, because it has the desirable property of allowing assignment of responsibility for fixing disparities that are discovered in the organization's data.

The proposal:

1. Appoint a chief diversity officer (CDO)[11] who is more than a figurehead.
2. Locate the CDO near the top of the organization's administrative hierarchy.
3. Give the CDO unrestricted access to data on outcomes received by the organization's staff and by its service recipients.
4. Provide the CDO with staff competent to conduct data analyses that can identify previously unrecognized demographic disparities, to determine whether those disparities are discriminatory or not, and to track the disparities over time.
5. Give the CDO authority to recommend new policies or practices throughout the organization.
6. Make the CDO's performance evaluation dependent on success in identifying and monitoring previously unrecognized disparities and in reducing documented discriminatory disparities.[12]

In the medical example that showed discriminatory disparities in diagnostic imaging for pediatric emergency patients (Marin et al., 2021), 14 physicians identified discriminatory disparities in a database that included treatment records from emergency departments of 44 hospitals over a 4-year period. Although these observed disparities have a plausible fix (i.e., the development of an evidence-based protocol to guide decisions on the use of diagnostic imaging), the physicians had no authority to impose that solution on emergency departments in the participating hospitals. The disparity-finding example for California police departments (see Box 1) was likewise the result of effort by persons in California's Department of Justice who had no authority to impose solution strategies on the police departments that produced the discriminatory disparities. These problems of disconnects between disparity finders and decision makers with authority to implement fixes for the disparities should not exist with the suggested arrangement of a CDO who can supervise both the disparity-finding effort and the implementation of fixes for discovered discriminatory disparities. For example, in the policing case, it might be wise to decentralize disparity finding so that it is done within each police department.

A six-item organizational self-test

1. Does your organization have data that allow determination of whether its employees are receiving equal treatment?
2. Does your organization have data that allow determination of whether those to whom it provides services are receiving equal treatment?[13]
3. Does your organization have someone with sufficient data-analysis skills to identify existing disparities and determine whether they are discriminatory?
4. Does your organization have an officer who has oversight for all DEI activities—someone who would know enough about your organization to answer the three preceding questions?
5. Has your organization ever identified a previously unrecognized discriminatory disparity?
6. Has your organization ever followed up on evidence for a discriminatory disparity by (a) implementing fixes expected to eliminate that disparity and (b) determining the extent to which the disparity was eliminated?

An organization that can answer yes to all of the first four questions can be judged well positioned to deal with DEI concerns. An organization that can answer yes to all six questions might be judged worthy of an award.

Appendix A

"Standard" (7-block) Implicit Association Test procedure

As most frequently used in research, an Implicit Association Test (IAT) consists of seven sets (blocks) of trials in which stimuli from four categories are classified. Any IAT is completely specified by the labels to be used for the four categories and the stimulus items (exemplars) used to represent each of the four categories. The subject's task in each of the seven blocks is to provide correct classifications of stimulus items (generally by pressing an assigned left- or right-positioned key on a computer keyboard—e.g., "E" and "I" on a QWERTY keyboard) into their categories. Typically, two of the categories are called *target* categories. The first reported IAT (Experiment 1 in Greenwald et al., 1998) used *flower* and *insect* as the labels for its two target categories. The other two categories are *attribute* categories. These were *pleasant* and *unpleasant* (valence) in the flower–insect attitude IAT.

The standard order of seven blocks (typical trial numbers [totaling 190] in parentheses) is as follows:

1. Classify the items for the two target categories (20 trials)
2. Classify the items for the two attribute categories (20 trials)
3. Classify items for all four categories, one attribute and one target category assigned to each of the two keys, using the same assignment of categories to left and right keys as in Blocks 1 and 2 (20 trials)
4. Same as Block 3 (40 trials)

5. Classify the two target categories, reversing the key assignments of Block 1 and having more trials than in Block 1 (30 trials)
6. Classify items for all four categories, using the same reversed key assignments of the target categories as in Block 5 (20 trials)
7. Same as Block 6 (40 trials)

The IAT is often administered with computer software that records latency to occurrence of the correct response, recording occurrence of error responses but not registering the trial's latency as completed until the correct response occurs. The usefulness of this method was demonstrated by Greenwald et al. (2003). When the IAT is administered with a procedure that records latency to the first response on a trial, if that response is an error, an error penalty is added to that trial in recording the latency for computing the *D* measure.

Transparency

Editor: Nora S. Newcombe
Declaration of Conflicting Interests
The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

Acknowledgments

The authors are grateful to multiple colleagues for their contributions to a prior manuscript titled "The Implicit Association Test at Age 20: What Is Known and What Is Not Known About Implicit Bias" (Greenwald et al., 2020). The material in this article's section headed What Is Known About Implicit Bias was in large part condensed from more detailed treatments in that predecessor, which also spun off an article (on best practices for using the Implicit Association Test in research) that appeared recently in *Behavior Research Methods* (Greenwald et al., 2021). Those contributors are Miguel Brendl, Huajian Cai, Dario Cvenček, Adam Hahn, Eric Hehman, Wilhelm Hofmann, Sean Joseph Hughes, Ian Hussey, Christian Jordan, John Jost, Teri A. Kirby, Calvin K. Lai, Jonas W. B. Lang, Kristen P. Lindgren, Dominika Maison, Brian D. Ostafin, James R. Rae, Kate A. Ratliff, Colin T. Smith, Adriaan Spruyt, Reinout W. Wiers, Mahzarin R. Banaji, Yoav Bar-Anan, Tessa E. S. Charlesworth, Jan De Houwer, John F. Kihlstrom, Benedek Kurdi, Franziska Meissner, Greg Mitchell, Brian A. Nosek, Marco Perugini, Klaus Rothermund, and Jeffrey Sherman. Other colleagues who provided valuable advice for this article are Aylin Caliskan, Sapna Cheryan, Lorie Fridell, Jack Glaser, Phillip Atiba Goff, Richard Gonzalez, Cheryl Kaiser, Janice Sabin, and Claude M. Steele.

Notes

1. Readers unfamiliar with the IAT will find greater description of it in the later section What Is Known About Implicit Bias.
2. This decline was documented in data from major U.S. national surveys by Schuman et al. (1997; an overview of their findings is available in Appendix 1 of Banaji & Greenwald, 2013, esp. pp. 175–179).
3. Banaji and Greenwald (2013, p. 158) analyzed data from more than 1.5 million persons who had completed both self-report race-attitude measures and parallel IAT measures on the Project Implicit website. Forty percent of these persons showed a combination of egalitarian attitude (no race preference) on self-report measures and more-than-minor implicit (IAT) racial preference for White relative to Black.
4. Findings described here are covered in greater detail by Greenwald et al. (2020).
5. Findings of Rudman, Greenwald, and McGhee's experiments were first partially reported in 1996 (conference presentation cited in Greenwald et al., 1998, p. 1478). A second preliminary report appeared in the published proceedings of a 1998 conference that included a preliminary presentation of balanced identity theory (Greenwald et al., 2000).
6. A brief history of the entry of the priming concept into social psychology can be found in Greenwald and Banaji (2017, pp. 863–864). Very informative treatments of the subsequent development of priming research in social psychology can be found in Higgins et al. (1977) and Higgins and Bargh (1987). The understanding that priming procedures produce ephemeral effects that are qualitatively different from durable alterations in mental representations was treated by (among others) Higgins et al. (1985).
7. For the IATs in Stone et al.'s experiment, the contrasted racial categories were operationally defined using "pictures of three White American men and three Hispanic American men" (p. 97).
8. Extending data collection in this fashion to increase power in testing a result that is statistically marginal on an initial analysis is problematic—in the sense of producing publication bias—only if researchers selectively choose not to report findings when the enlarged sample yields a nonsignificant result.
9. A concern about the self–anxiety IATs used in clinical studies (see the review by Roefs et al., 2011) is due to the valence difference between calmness (positive) and anxiety (negative), which likely allows these IATs to pick up implicit self-esteem in addition to the desired implicit self–anxiety association (see Greenwald et al., 2021).
10. An example is useful. Suppose that, in a large company, Black occupants of an entry-level management position are less likely to be promoted to the next higher level than are White occupants of the same position. Higher management may say that the difference is not a discriminatory disparity because the White occupants have higher performance ratings from their supervisors. But those differences in performance ratings could, themselves, be discriminatory disparities. The nondiscriminatory interpretation could be advanced more appropriately by finding differences in objective indicators of the entry-level managers' performance (e.g., sales performance of their supervisees, turnover among their supervisees, punctuality of their supervisees, etc.).
11. Large organizations will have subunits across which criteria for DEI might appropriately vary and for which the workload of managing DEI justifies horizontal distribution of authority (subsidiary to the CDO) to those subunits. Still, identifying a single CDO with oversight responsibility for all of those subunits is desirable because it gives the organization's top leadership (i.e.,

the CDO's immediate supervisors) ultimate responsibility for the organization's effectiveness in DEI.

12. Should the CDO's efforts be overridden by some other actor with appropriate authority, obviously responsibility for success or failure of corrective efforts belongs to that other actor.

13. We understand that equitable treatment is a more appropriate policy goal than equal treatment. The first two self-test questions are nevertheless stated in terms of equal treatment, because disparity finding can reveal inequitable treatment (a "discriminatory disparity") only by first observing a previously unrecognized inequality of outcomes, then determining that the inequality has no plausible nondiscriminatory explanation.

References

Adler, A. D., Strunk, D. R., & Fazio, R. H. (2015). What changes in cognitive therapy for depression? An examination of cognitive therapy skills and maladaptive beliefs. *Behavior Therapy*, *46*(1), 96–109. https://doi.org/10.1016/j.beth.2014.09.001

Agency for Healthcare Research and Quality. (2020). *2019 National Healthcare Quality and Disparities Report* (Publication No. 20(21)-0045-EF). U.S. Department of Health and Human Services. https://www.ahrq.gov/sites/default/files/wysiwyg/research/findings/nhqrdr/2019qdr-final-es-cs061721.pdf

American Public Health Association. (n.d.). *Racism and health*. Retrieved January 4, 2022, from https://www.apha.org/Topics-and-Issues/Health-Equity/Racism-and-health

American Public Health Association. (2018, November 3). *Achieving health equity in the United States* (APHA Policy No. 20189). https://www.apha.org/policies-and-advocacy/public-health-policy-statements/policy-database/2019/01/29/achieving-health-equity

Asendorpf, J. B., Banse, R., & Mücke, D. (2002). Double dissociation between implicit and explicit personality self-concept: The case of shy behavior. *Journal of Personality and Social Psychology*, *83*, 380–393. https://doi.org/10.1037/0022-3514.83.2.380

Banaji, M. R., & Greenwald, A. G. (2013). *Blindspot: Hidden biases of good people*. Delacorte Press.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610–623). Association for Computing Machinery. https://doi.org/10.1145/3442188.3445922

Blair, I. V. (2002). The malleability of automatic stereotypes and prejudice. *Personality and Social Psychology Review*, *6*, 242–261. https://doi.org/10.1207/S15327957PSPR0603_8

Blanton, H., & Jaccard, J. (2006). Arbitrary metrics in psychology. *American Psychologist*, *61*(1), 27–41. https://doi.org/10.1037/0003-066X.61.1.27

Blanton, H., Jaccard, J., Strauts, E., Mitchell, G., & Tetlock, P. E. (2015). Toward a meaningful metric of implicit prejudice. *Journal of Applied Psychology*, *100*, 1468–1481. https://doi.org/10.1037/a0038379

Bolukbasi, T., Chang, K. -W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In D. Lee, M. Sugiyama, U. von Luxburg, I. Guyon, & R. Garnett (Eds.), *NIPS'16: Proceedings of the 30th International Conference on Neural Information Processing Systems* (pp. 4356–4364). Neural Information Processing Systems. http://papers.nips.cc/paper/6228-man-is-to-computer-programmer-as-woman-is-to-homemaker-debiasing-word-embeddings.pdf

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Wiley. https://doi.org/10.1002/9780470743386

Brendl, C. M., Markman, A. B., & Messner, C. (2001). How do indirect measures of evaluation work? Evaluating the inference of prejudice in the implicit association test. *Journal of Personality and Social Psychology*, *81*, 760–773. https://doi.org/10.1037/0022-3514.81.5.760

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183–186. https://doi.org/10.1126/science.aal4230

Cameron, C. D., Brown-Iannuzzi, J. L., & Payne, B. K. (2012). Sequential priming measures of implicit social cognition: A meta-analysis of associations with behavior and explicit attitudes. *Personality and Social Psychology Review*, *16*(4), 330–350. https://doi.org/10.1177/1088868312440047

Carnes, M., Devine, P. G., Manwell, L. B., Byars-Winston, A., Fine, E., Ford, C. E., Forscher, P., Isaac, C., Kaatz, A., Magua, W., Palta, M., & Sheridan, J. (2015). The effect of an intervention to break the gender bias habit for faculty at one institution: A cluster randomized, controlled trial. *Academic Medicine*, *90*(2), 221–230. https://doi.org/10.1097/ACM.0000000000000552

Chang, E. H., Milkman, K. L., Gromet, D. M., Rebele, R. W., Massey, C., Duckworth, A. L., & Grant, A. M. (2019). The mixed effects of online diversity training. *Proceedings of the National Academy of Sciences, USA*, *116*, 7778–7783. https://doi.org/10.1073/pnas.1816076116

Charbonneau, A., & Glaser, J. (2021). Suspicion and discretion in policing: How laws and policies contribute to inequity. *UC Irvine Law Review*, *11*(5), 1327–1348. https://scholarship.law.uci.edu/ucilr/vol11/iss5/6.

Civil Rights Act of 1964, Pub. L. No. 88-352, 78 Stat. 241 (1964). https://www.govinfo.gov/content/pkg/STATUTE-78/pdf/STATUTE-78-Pg241.pdf

Clermont, K. M., & Schwab, S. J. (2009). Employment discrimination plaintiffs in federal court: From bad to worse? *Harvard Law & Policy Review*, *3*, 103–132. https://harvardlpr.com/wp-content/uploads/sites/20/2013/05/3.1_6_Clermont.pdf

Connor, P., & Evers, E. R. K. (2020). The bias of individuals (in crowds): Why implicit bias is probably a noisily-measured individual-level construct. *Perspectives on Psychological Science*, *15*(6), 1329–1345.

Cone, J., Mann, T. C., & Ferguson, M. J. (2017). Changing our implicit minds: How, when, and why implicit evaluations can be rapidly revised. In J. M. Olson (Ed.), *Advances*

in experimental social psychology (Vol. 56, pp. 131–199). Academic Press.

Conrey, F. R., Sherman, J. W., Gawronski, B., Hugenberg, K., & Groom, C. J. (2005). Separating multiple processes in implicit social cognition: The Quad model of implicit task performance. *Journal of Personality and Social Psychology*, *89*, 469–487. https://doi.org/10.1037/0022-3514.89.4.469

Cristea, I. A., & Kok, R. N., & Cuijpers, P. (2016). The effectiveness of cognitive bias modification interventions for substance addictions: A meta-analysis. *PLOS ONE*, *11*(9), Article e0162226. https://doi.org/10.1371/journal.pone.0162226

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*, 281–302. https://doi.org/10.1037/h0040957

Cvencek, D., Greenwald, A. G., McLaughlin, K. A., & Meltzoff, A. N. (2020). Early implicit-explicit discrepancies in self-esteem as correlates of childhood depressive symptoms. *Journal of Experimental Child Psychology*, *200*, Article 104962. https://doi.org/10.1016/j.jecp.2020.104962

Cvencek, D., Greenwald, A. G., & Meltzoff, A. N. (2011). Measuring implicit attitudes of 4-year-old children: The Preschool Implicit Association Test. *Journal of Experimental Child Psychology*, *109*, 187–200. https://doi.org/10.1016/j.jecp.2010.11.002

Cvencek, D., Meltzoff, A. N., Maddox, C. D., Nosek, B. A., Rudman, L. A., Devos, T., Dunham, Y., Baron, A. S., Steffens, M. C., Lane, K., Horcajo, J., Ashburn-Nardo, L., Quinby, A., Srivastava, S. B., Schmidt, K., Aidman, E., Tang, E., Farnham, S., Mellott, D. S., . . . Greenwald, A. G. (2021). Meta-analytic use of balanced identity theory to validate the Implicit Association Test. *Personality and Social Psychology Bulletin*, *47*, 185–200. https://doi.org/10.1177/0146167220916631

Dasgupta, N., & Asgari, S. (2004). Seeing is believing: Exposure to counterstereotypic women leaders and its effect on the malleability of automatic gender stereotyping. *Journal of Experimental Social Psychology*, *40*(5), 642–658. https://doi.org/10.1016/j.jesp.2004.02.003

Dasgupta, N., & Greenwald, A. G. (2001). On the malleability of automatic attitudes: Combating automatic prejudice with images of admired and disliked individuals. *Journal of Personality and Social Psychology*, *81*, 800–814. https://doi.org/10.1037/0022-3514.81.5.800

Dennehy, T. C., & Dasgupta, N. (2017). Female peer mentors early in college increase women's positive academic experiences and retention in engineering. *Proceedings of the National Academy of Sciences, USA*, *114*(23), 5964–5969. https://doi.org/10.1073/pnas.1613117114

Devine, P. G., Forscher, P. S., Austin, A. J., & Cox, W. T. L. (2012). Long-term reduction in implicit race bias: A prejudice habit-breaking intervention. *Journal of Experimental Social Psychology*, *48*, 1267–1278. https://doi.org/10.1016/j.jesp.2012.06.003

Devine, P. G., Forscher, P. S., Cox, W. T. L., Kaatz, A., Sheridan, J., & Carnes, M. (2017). A gender bias habit-breaking intervention led to increased hiring of female faculty in STEMM departments. *Journal of Experimental Social Psychology*, *73*, 211–215. https://doi.org/10.1016/j.jesp.2017.07.002

Dobbin, F., & Kalev, A. (2013). The origins and effects of corporate diversity programs. In Q. M. Roberson (Ed.), *The Oxford handbook of diversity and work* (pp. 253–281). Oxford University Press.

Dovidio, J. F., Kawakami, K., & Gaertner, S. L. (2002). Implicit and explicit prejudice and interracial interaction. *Journal of Personality and Social Psychology*, *82*, 62–68. https://doi.org/10.1037/0022-3514.82.1.62

Epidemiology. (2021, December 29). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Epidemiology&oldid=1062531057

Epstein, R. A. (2005). Disparities and discrimination in health care coverage: A critique of the Institute of Medicine study. *Perspectives in Biology and Medicine*, *48*(1Suppl), S26–S41. https://doi.org/10.1353/pbm.2005.0037

Fazio, R. H. (1990). Multiple processes by which attitudes guide behavior: The MODE model as an integrative framework. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. *23*, pp. 75–109). Academic Press. https://doi.org/10.1016/S0065-2601(08)60318-4

Fazio, R. H., Sanbonmatsu, D. M., Powell, M. C., & Kardes, F. R. (1986). On the automatic activation of attitudes. *Journal of Personality and Social Psychology*, *50*, 229–238. https://doi.org/10.1037/0022-3514.50.2.229

Forscher, P. S., Lai, C. K., Axt, J. R., Ebersole, C. R., Herman, M., Devine, P. G., & Nosek, B. A. (2019). A meta-analysis of procedures to change implicit measures. *Journal of Personality and Social Psychology*, *117*(3), 522–559. https://doi.org/10.1037/pspa0000160

Forscher, P. S., Mitamura, C., Dix, E. L., Cox, W. T. L., & Devine, P. G. (2017). Breaking the prejudice habit: Mechanisms, timecourse, and longevity. *Journal of Experimental Social Psychology*, *72*, 133–146. https://doi.org/10.1016/j.jesp.2017.04.009

Gamer, J., Schmukle, S. C., Luka-Krausgrill, U., & Egloff, B. (2008). Examining the dynamics of the implicit and the explicit self-concept in social anxiety: Changes in the Implicit Association Test-Anxiety and the Social Phobia Anxiety Inventory following treatment. *Journal of Personality Assessment*, *90*(5), 476–480. https://doi.org/10.1080/00223890802248786

Gawronski, B., Morrison, M., Phills, C. E., & Galdi, S. (2017). Temporal stability of implicit and explicit measures: A longitudinal analysis. *Personality and Social Psychology Bulletin*, *43*(3), 300–312. https://doi.org/10.1177/0146167216684131

Goldin, C., & Rouse, C. (2000). Orchestrating impartiality: The impact of "blind" auditions on female musicians. *American Economic Review*, *90*, 715–741. https://doi.org/10.1016/j.jesp.2017.04.009

Gonen, H., & Goldberg, Y. (2019). Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, *Volume 1 (Long and Short Papers)*

(pp. 609–614). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1061

Greenwald, A. G. (2009). What (and where) is the ethical code concerning researcher conflict of interest? *Perspectives on Psychological Science*, *4*, 32–35. https://doi.org/10.1111/j.1745-6924.2009.01086.x

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review*, *102*(1), 4–27. https://doi.org/10.1037/0033-295X.102.1.4

Greenwald, A. G., & Banaji, M. R. (2017). The implicit revolution: Reconceiving the relation between conscious and unconscious. *American Psychologist*, *72*(9), 861–871. https://doi.org/10.1037/amp0000238

Greenwald, A. G., Banaji, M. R., & Nosek, B. A. (2015). Statistically small effects of the Implicit Association Test can have societally large effects. *Journal of Personality and Social Psychology*, *108*, 553–561. https://doi.org/10.1037/pspa0000016

Greenwald, A. G., Banaji, M. R., Rudman, L. A., Farnham, S. D., Nosek, B. A., & Mellott, D. S. (2002). A unified theory of implicit attitudes, stereotypes, self-esteem, and self-concept. *Psychological Review*, *109*, 3–25. https://doi.org/10.1037/0033-295X.109.1.3

Greenwald, A. G., Banaji, M. R., Rudman, L. A., Farnham, S. D., Nosek, B. A., & Rosier, M. (2000). Prologue to a unified theory of attitudes, stereotypes, and self-concept. In J. P. Forgas (Ed.), *Feeling and thinking: The role of affect in social cognition* (pp. 308–330). Cambridge University Press.

Greenwald, A. G., Brendl, M., Cai, H., Charlesworth, T., Cvencek, D., Dovidio, J. F., Friese, M., Hahn, A., Hehman, E., Hofmann, W., Hughes, S., Hussey, I., Jordan, C., Jost, J., Kirby, T., Lai, C. K., Lang, J., Lindgren, K. P., Maison, D., . . . Wiers, R. W. (2020). *The Implicit Association Test at age 20: What is known and what is not known about implicit bias*. PsyArXiv. https://doi.org/10.31234/osf.io/bf97c

Greenwald, A. G., Brendl, M., Cai, H., Cvencek, D., Dovidio, J. F., Friese, M., Hahn, A., Hehman, E., Hofmann, W., Hughes, S., Hussey, I., Jordan, C., Kirby, T. A., Lai, C. K., Lang, J., Lindgren, K. P., Maison, D., Ostafin, B. D., Rae, J. R., . . . Wiers, R. W. (2021). Best research practices for using the Implicit Association Test. *Behavior Research Methods*. Advance online publication. https://doi.org/10.3758/s13428-021-01624-3

Greenwald, A. G., & Lai, C. K. (2020). Implicit social cognition. *Annual Review of Psychology*, *71*, 419–445. https://doi.org/10.1146/annurev-psych-010419-050837

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The Implicit Association Test. *Journal of Personality and Social Psychology*, *74*, 1464–1480. https://doi.org/10.1037/0022-3514.74.6.1464

Greenwald, A. G., Nosek, B. A., & Banaji, M. R. (2003). Understanding and using the Implicit Association Test: I. An improved scoring algorithm. *Journal of Personality and Social Psychology*, *85*, 197–216. https://doi.org/10.1037/0022-3514.85.2.197

Greenwald, A. G., & Pettigrew, T. F. (2014). With malice toward none and charity for some: Ingroup favoritism enables discrimination. *American Psychologist*, *69*, 669–684. https://doi.org/10.1037/a0036056

Greenwald, A. G., Poehlman, T. A., Uhlmann, E. L., & Banaji, M. R. (2009). Understanding and using the Implicit Association Test: III. Meta-analysis of predictive validity. *Journal of Personality and Social Psychology*, *97*, 17–41. https://doi.org/10.1037/a0015575

Gregg, A. P., Seibt, B., & Banaji, M. R. (2006). Easier done than undone: Asymmetry in the malleability of implicit preferences. *Journal of Personality and Social Psychology*, *90*(1), 1–20. https://doi.org/10.1037/0022-3514.90.1.1

Hahn, A., Judd, C. M., Hirsh, H. K., & Blair, I. V. (2014). Awareness of implicit attitudes. *Journal of Experimental Psychology: General*, *143*(3), 1369–1392. https://doi.org/10.1037/a0035028

Hailey, S. E., & Olson, K. R. (2013). A social psychologist's guide to the development of racial attitudes. *Social and Personality Psychology Compass*, *7*(7), 457–469. https://doi.org/10.1111/spc3.12038

Hart, M. (2005). Subjective decision-making and unconscious discrimination. *Alabama Law Review*, *56*, 741–791.

Heider, F. (1958). *The psychology of interpersonal relations*. Wiley. https://doi.org/10.1037/10628-000

Heilman, M. E., & Haynes, M. C. (2008). Subjectivity in the appraisal process: A facilitator of gender bias in work settings. In E. Borgida & S.T. Fiske (Eds.), *Psychological science in the courtroom* (pp. 127–156). Blackwell. https://doi.org/10.1002/9780470696422.ch7

Higgins, E. T., & Bargh, J. A. (1987). Social cognition and social perception. *Annual Review of Psychology*, *38*, 369–425. https://doi.org/10.1146/annurev.ps.38.020187.002101

Higgins, E. T., Bargh, J. A., & Lombardi, W. J. (1985). Nature of priming effects on categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *11*, 59–69. https://doi.org/10.1037/0278-7393.11.1.59

Higgins, E. T., Rholes, W. S., & Jones, C. R. (1977). Category accessibility and impression formation. *Journal of Experimental Social Psychology*, *13*, 141–154. https://doi.org/10.1016/S0022-1031(77)80007-3

Hofmann, W., Gawronski, B., Gschwendner, T., Le, H., & Schmitt, M. (2005). A meta-analysis on the correlation between the Implicit Association Test and explicit self-report measures. *Personality and Social Psychology Bulletin*, *31*, 1369–1385. https://doi.org/10.1177/0146167205275613

Huijding, J., & de Jong, P. J. (2007). Beyond fear and disgust: The role of (automatic) contamination-related associations in spider phobia. *Journal of Behavior Therapy and Experimental Psychiatry*, *38*(2), 200–211. https://doi.org/10.1016/j.jbtep.2006.10.009

Ito, T. A., Friedman, N. P., Bartholow, B. D., Correll, J., Loersch, C., Altamirano, L. J., & Miyake, A. (2015). Toward a comprehensive understanding of executive cognitive function in implicit racial bias. *Journal of Personality and Social Psychology*, *108*(2), 187–218. https://doi.org/10.1037/a0038557

Johnson, L. B. (1964, June 4). *Commencement Address at Howard University: "To Fulfill These Rights."* Teaching American History. https://teachingamericanhistory.org/document/commencement-address-at-howard-university-to-fulfill-these-rights/

Kahneman, D., & Klein, G. (2009). Conditions for intuitive expertise: A failure to disagree. *American Psychologist*, *64*(6), 515–526. https://doi.org/10.1037/a0016755

Kaiser, C. R., Major, B., Jurcevic, I., Dover, T. L., Brady, L. M., & Shapiro, J. R. (2013). Presumed fair: Ironic effects of organizational diversity structures. *Journal of Personality and Social Psychology*, *104*(3), 504–519. https://doi.org/10.1037/a0030838

Kalev, A., Dobbin, F., & Kelly, E. (2006). Best practices or best guesses? Assessing the efficacy of corporate affirmative action and diversity policies. *American Sociological Review*, *71*, 589–617. https://doi.org/10.1177/000312240607100404

Klauer, K. C., Schmitz, F., Teige-Mocigemba, S., & Voss, A. (2010). Understanding the role of executive control in the implicit association test: Why flexible people have small IAT effects. *The Quarterly Journal of Experimental Psychology*, *63*(3), 595–619. https://doi.org/10.1080/17470210903076826

Kurdi, B., Seitchik, A. E., Axt, J. R., Carroll, T. J., Karapetyan, A., Kaushik, N., Tomezsko, D., Greenwald, A. G., & Banaji, M. R. (2019). Relationship between the Implicit Association Test and intergroup behavior: A meta-analysis. *American Psychologist*, *74*(5), 569–586. https://doi.org/10.1037/amp0000364

Lai, C. K., Marini, M., Lehr, S. A., Cerruti, C., Shin, J.-E. L., Joy-Gaba, J. A., Ho, A. K., Teachman, B. A., Wojcik, S. P., Koleva, S. P., Frazier, R. S., Heiphetz, L., Chen, E. E., Turner, R. N., Haidt, J., Kesebir, S., Hawkins, C. B., Schaefer, H. S., Rubichi, S., . . . Nosek, B. A. (2014). Reducing implicit racial preferences: I. A comparative investigation of 17 interventions. *Journal of Experimental Psychology: General*, *143*(4), 1765–1785. https://doi.org/10.1037/a0036260

Lai, C. K., Skinner, A. L., Cooley, E., Murrar, S., Brauer, M., Devos, T., Calanchini, J., Xiao, Y. J., Pedram, C., Marshburn, C. K., Simon, S., Blanchar, J. C., Joy-Gaba, J. A., Conway, J., Redford, L., Klein, R. A., Roussos, G., Schellhaas, F. M. H., Burns, M., . . . Nosek, B. A. (2016). Reducing implicit racial preferences: II. Intervention effectiveness across time. *Journal of Experimental Psychology: General*, *145*, 1001–1016. https://doi.org/10.1037/xge0000179

Leadership Conference on Civil and Human Rights. (2011). *Restoring a national consensus: The need to end racial profiling in America*. https://wiki.duke.edu/download/attachments/100729513/racial_profiling2011.pdf

Leslie, L. M. (2019). Diversity initiative effectiveness: A typological theory of unintended consequences. *Academy of Management Review*, *44*, 538–563. https://doi.org/10.5465/amr.2017.0087

Lindgren, K. P., Wiers, R. W., Teachman, B. A., Gasser, M. L., Westgate, E. C., Cousijn, J., Enkema, M. C., & Neighbors, C. (2015). Attempted training of alcohol approach and drinking identity associations in US undergraduate drinkers: Null results from two studies. *PLOS ONE*, *10*(8), Article e0134642. https://doi.org/10.1371/journal.pone.0134642

Mann, T. C., Kurdi, B., & Banaji, M. R. (2020). How effectively can implicit evaluations be updated? Using evaluative statements after aversive repeated evaluative pairings. *Journal of Experimental Psychology: General*, *149*(6), 1169–1192. https://doi.org/10.1037/xge0000701 https://doi.org/10.1037/xge0000701

Marin, J. R., Rodean, J., Hall, M., Alpern, E. R., Aronson, P. L., Chaudhari, P. P., Cohen, E., Freedman, S. B., Morse, R. B., Peltz, A., Samuels-Kalow, M., Shah, S. S., Simon, H. K., & Neuman, M. I. (2021). Racial and ethnic differences in emergency department diagnostic imaging at US Children's Hospitals, 2016-2019. *JAMA Network Open*, *4*(1), Article e2033710. https://doi.org/10.1001/jamanetworkopen.2020.33710

McCarthy, J. M., Van Iddekinge, C. H., & Campion, M. A. (2010). Are highly structured job interviews resistant to demographic similarity effects? *Personnel Psychology*, *63*, 325–359. https://doi.org/10.1111/j.1744-6570.2010.01172.x

Meissner, F., & Rothermund, K. (2013). Estimating the contributions of associations and recoding in the Implicit Association Test: The ReAL model for the IAT. *Journal of Personality and Social Psychology*, *104*(1), 45–69. https://doi.org/10.1037/a0030734

Mierke, J., & Klauer, K. C. (2001). Implicit association measurement with the IAT: Evidence for effects of executive control processes. *Zeitschrift für Experimentelle Psychologie*, *48*(2), 107–122. https://doi.org/10.1026//0949-3946.48.2.107

Mor Barak, M. E., Lizano, E. L., Kim, A., Duan, L., Rhee, M. K., Hsiao, H. Y., & Brimhall, K. C. (2016). The promise of diversity management for climate of inclusion: A state-of-the-art review and meta-analysis. *Human Service Organizations: Management, Leadership & Governance*, *40*, 305–333. https://doi.org/10.1080/23303131.2016.1138915

Nishii, L. H., Khattab, J., Shemla, M., & Paluch, R. M. (2018). A multi-level process model for understanding diversity practice effectiveness. *Academy of Management Annals*, *12*, 37–82. https://doi.org/10.5465/annals.2016.0044

Nosek, B. A. (2005). Moderators of the relationship between implicit and explicit evaluation. *Journal of Experimental Psychology: General*, *134*(4), 565–584. https://doi.org/10.1037/0096-3445.134.4.565

Nosek, B. A., & Hansen, J. J. (2008). The associations in our heads belong to us: Searching for attitudes and knowledge in implicit evaluation. *Cognition and Emotion*, *22*, 553–594. https://doi.org/10.1080/02699930701438186

Nosek, B. A., Smyth, F. L., Hansen, J. J., Devos, T., Lindner, N. M., Ranganath, K. A., Smith, C. T., Olson, K. R., Chugh, D., Greenwald, A. G., & Banaji, M. (2007). Pervasiveness and correlates of implicit attitudes and stereotypes. *European Review of Social Psychology*, *18*, 36–88. https://doi.org/10.1080/10463280701489053

Onyeador, I. N., Wittlin, N. M., Burke, S. E., Dovidio, J. F., Perry, S. P., Hardeman, R. R., Dyrbye, L. N., Herrin, J., Phelan, S. M., & van Ryn, M. (2020). The value of interracial contact for reducing anti-Black bias among non-Black

physicians: A Cognitive Habits and Growth Evaluation (CHANGE) Study report. *Psychological Science*, *31*(1), 18–30. https://doi.org/10.1177/0956797619879139

Osgood, C. E., & Tannenbaum, P. H. (1955). The principle of congruity in the prediction of attitude change. *Psychological Review*, *62*, 42–55. https://doi.org/10.1037/h0048153

Oswald, F. L., Mitchell, G., Blanton, H., Jaccard, J., & Tetlock, P. E. (2013). Predicting ethnic and racial discrimination: A meta-analysis of IAT criterion studies. *Journal of Personality and Social Psychology*, *105*, 171–192. http://doi.org/10.1037/a0032734

Oswald, F. L., Mitchell, G., Blanton, H., Jaccard, J., & Tetlock, P. E. (2015). Using the IAT to predict ethnic and racial discrimination: Small effect sizes of unknown societal significance. *Journal of Personality and Social Psychology*, *108*(4), 562–571. https://doi.org/10.1037/pspa0000023

Ottaway, S. A., Hayden, D. C., & Oakes, M. A. (2001). Implicit attitudes and racism: Effects of word familiarity and frequency on the implicit association test. *Social Cognition*, *19*(2), 97–144. https://doi.org/10.1521/soco.19.2.97.20706

Paine, L., de la Rocha, P., Eyssallenne, A. P., Andrews, C. A., Loo, L., Jones, C. P., Collins, A. M., & Morse, M. (2021). Declaring racism a public health crisis in the United States: Cure, poison, or both? *Frontiers in Public Health*, *9*, Article 676784. https://doi.org/10.3389/fpubh.2021.676784

Paluck, E. L., & Green, D. P. (2009). Prejudice reduction: What works? A review and assessment of research and practice. *Annual Review of Psychology*, *60*, 339–367. https://doi.org/10.1146/annurev.psych.60.110707.163607

Paluck, E. L., Porat, R., Clark, C. S., & Green, D. P. (2021). Prejudice reduction: Progress and challenges. *Annual Review of Psychology*, *72*, 533–560. https://doi.org/10.1146/annurev-psych-071620-030619

Payne, B. K., & Gawronski, B. (2010). A history of implicit social cognition: Where is it coming from? Where is it now? Where is it going? In B. Gawronski & B. K. Payne (Eds.), *Handbook of implicit social cognition: Measurement, theory, and applications* (pp. 1–15). The Guilford Press.

Payne, B. K., Vuletich, H. A., & Lundberg, K. B. (2017). The bias of crowds: How implicit bias bridges personal and systemic prejudice. *Psychological Inquiry*, *28*(4), 233–248. https://doi.org/10.1080/1047840X.2017.1335568

Pennington, C. R., Campbell, C., Monk, R. L., & Heim, D. (2016). The malleability of stigmatizing attitudes: Combining imagined social contact with implicit attitude feedback. *American Journal of Psychiatric Rehabilitation*, *19*(3), 175–195. https://doi.org/10.1080/15487768.2016.1171175

Pettigrew, T. F., & Tropp, L. R. (2006). A meta-analytic test of intergroup contact theory. *Journal of Personality and Social Psychology*, *90*(5), 751–783. https://doi.org/10.1037/0022-3514.90.5.751

Pew Research Center. (2019). *Race in America 2019*. https://www.pewresearch.org/social-trends/2019/04/09/race-in-america-2019/

Racial and Identity Profiling Advisory Board. (2021a). *Appendices to annual report 2021*. https://oag.ca.gov/sites/all/files/agweb/pdfs/ripa/ripa-appendices-2021.pdf

Racial and Identity Profiling Advisory Board. (2021b). *Racial and Identity Profiling Advisory Board: 2021 report quick facts*. https://oag.ca.gov/sites/all/files/agweb/pdfs/ripa/ripa-quick-facts-2021-01.pdf?

Racial and Identity Profiling Advisory Board. (2021c). *Racial and Identity Profiling Advisory Board: Annual report 2021*. https://oag.ca.gov/sites/all/files/agweb/pdfs/ripa/ripa-board-report-2021.pdf

Renna, M. E., Seeley, S. H., Heimberg, R. G., Etkin, A., Fresco, D. M., & Mennin, D. S. (2018). Increased attention regulation from emotion regulation therapy for generalized anxiety disorder. *Cognitive Therapy Resource*, *42*, 121–134. https://doi-org.proxy.lib.duke.edu/10.1007/s10608-017-9872-7

Roefs, A., Huijding, J., Smulders, F. T. Y., MacLeod, C. M., de Jong, P. J., Wiers, R. W., & Jansen, A. T. M. (2011). Implicit measures of association in psychopathology research. *Psychological Bulletin*, *137*(1), 149–193. https://doi.org/10.1037/a0021729

Rothermund, K., Teige-Mocigemba, S., Gast, A., & Wentura, D. (2009). Minimizing the influence of recoding in the Implicit Association Test: The Recoding-Free Implicit Association Test (IAT-RF). *The Quarterly Journal of Experimental Psychology*, *62*, 84–98. https://doi.org/10.1080/17470210701822975

Rothermund, K., & Wentura, D. (2001). Figure-ground asymmetries in the Implicit Association Test (IAT). *Zeitschrift für Experimentelle Psychologie*, *48*, 94–106. https://doi.org/10.1026//0949-3946.48.2.94

Rothermund, K., & Wentura, D. (2004). Underlying processes in the Implicit Association Test: Dissociating salience from associations. *Journal of Experimental Psychology: General*, *133*, 139–165. https://doi.org/10.1037/0096-3445.133.2.139

Rudman, L. A., Greenwald, A. G., & McGhee, D. E. (2001). Implicit self-concept and evaluative implicit gender stereotypes: Self and ingroup share desirable traits. *Personality and Social Psychology Bulletin*, *27*(9), 1164–1178. https://doi.org/10.1177/0146167201279009

Schuhl, J., Lambert, E., & Chatard, A. (2019). Can imagination reduce prejudice over time? A preregistered test of the imagined contact hypothesis. *Basic and Applied Social Psychology*, *41*, 122–131. https://doi.org/10.1080/01973533.2019.1579719

Schuman, H., Steeh, C., Bobo, L., & Krysan, M. (1997). *Racial attitudes in America*. Harvard University Press.

Shook, N. J., & Fazio, R. H. (2008). Interracial roommate relationships: An experimental field test of the contact hypothesis. *Psychological Science*, *19*, 717–723. https://doi.org/10.1111/j.1467-9280.2008.02147.x

Steele, C. M. (1997). A threat in the air: How stereotypes shape intellectual identity and performance. *American Psychologist*, *52*, 613–629. https://doi.org/10.1037/0003-066X.52.6.613

Stone, J., Moskowitz, G. B., Zestcott, C. A., & Wolsiefer, K. J. (2020). Testing active learning workshops for reducing implicit stereotyping of Hispanics by majority and minority group medical students. *Stigma and Health*, *5*(1), 94–103. https://doi.org/10.1037/sah0000179

Stout, J. G., Dasgupta, N., Hunsinger, M., & McManus, M. A. (2011). STEMing the tide: Using ingroup experts to inoculate women's self-concept in science, technology, engineering, and mathematics (STEM). *Journal of Personality and Social Psychology*, *100*(2), 255–270. https://doi.org/10.1037/a0021385

Teachman, B. A., Marker, C. D., & Smith-Janik, S. B. (2008). Automatic associations and panic disorder: Trajectories of change over the course of treatment. *Journal of Consulting and Clinical Psychology*, *76*(6), 988–1002. https://doi.org/10.1037/a0013113

Teachman, B. A., & Woody, S. R. (2003). Automatic processing in spider phobia: Implicit fear associations over the course of treatment. *Journal of Abnormal Psychology*, *112*(1), 100–109. https://doi.org/10.1037/0021-843X.112.1.100

Turner, R. N., & Crisp, R. J. (2010). Imagining intergroup contact reduces implicit prejudice. *British Journal of Social Psychology*, *49*(1), 129–142. https://doi.org/10.1348/014466609X419901

Turner, R. N., Hewstone, M., & Voci, A. (2007). Reducing explicit and implicit outgroup prejudice via direct and extended contact: The mediating role of self-disclosure and intergroup anxiety. *Journal of Personality and Social Psychology*, *93*(3), 369–388. https://doi.org/10.1037/0022-3514.93.3.369

Vezzali, L., Capozza, D., Gionannini, S. D., & Stathi, S. (2012). Improving implicit and explicit intergroup attitudes using imagined contact: An experimental intervention with elementary school children. *Group Processes & Intergroup Relations*, *15*, 203–2012. https://doi.org/10.1177/1368430211424920

West, K., Hotchin, V., & Wood, C. (2017). Imagined contact can be more effective for participants with stronger initial prejudices. *Journal of Applied Social Psychology*, *47*, 282–292. https://doi.org/10.1111/jasp.12437

Wiers, R. W., Eberl, C., Rinck, M., Becker, E. S., & Lindenmeyer, J. (2011). Retraining automatic action tendencies changes alcoholic patients' approach bias for alcohol and improves treatment outcome. *Psychological Science*, *22*(4), 490–497. https://doi.org/10.1177/0956797611400615

Whitmer, G. (2020, August 5). *Executive directive 2020-09*. https://www.michigan.gov/whitmer/0,9309,7-387-90499_90704-535748--,00.html