Ergo AN OPEN ACCESS JOURNAL OF PHILOSOPHY

Do Implicit Racial Biases Have Significant Discriminatory Effects?

TIMOTHY J. FULLER

This article investigates whether implicit racial biases have significant discriminatory effects. To this end, it evaluates meta-analyses of studies on measures of implicit bias and behavioral effects to which they are correlated. On balance, I maintain, the best interpretation of these meta-analyses and relevant surrounding research supports the conclusion that implicit racial biases are significantly correlated to racially biased behaviors, with effects that are consequential at both the individual and group levels. This conclusion is compatible with, but does not entail, the proposition that implicit racial biases in fact cause such effects. In consequence, I consider the contribution implicit bias research might make to our best accounts of racial discrimination and inequality on both a casual and non-causal construal. I argue it is plausible that research on implicit racial bias, on either construal, will play a substantive role in such accounts.

1. Introduction

Implicit bias research is mired in controversy: Mitchell and Tetlock (2017) call it "wobbly science" (2017: 188); an editorial in *The Wall Street Journal* recently declared "its scientific basis is crumbling" (MacDonald 2017); and Edouard Machery (2017) has warned that building on its foundations may be like building on "quicksand." Grounds for skepticism about implicit bias research are varied. According to one set of criticisms, the most widespread test for implicit bias—the Implicit Association Test (IAT)—has problematic psychometric properties, including dubious methods for computing scores (Mitchell & Tetlock 2017), an arbitrary measurement scale (Blanton & Jaccard 2006), and low test-retest reliability (Machery 2016). Such criticisms will not be my focus here. Arguably, they do not support the strongest forms of skepticism about implicit bias research.

Contact: Timothy J. Fuller <timmy.fuller@gmail.com>

https://doi.org/10.3998/ergo.1159

Defenders of implicit bias research might maintain that despite its flaws, implicit measures such as the IAT are nevertheless imperfect tools that usefully predict biased behaviors which, at least in some circumstances, have significant discriminatory effects.

A second group of criticisms, arguably more formidable, questions whether implicit bias research has identified useful tools for predicting biased behaviors or discovered correlations and effect sizes that suggest implicit racial biases have significant discriminatory effects in real-world circumstances. One such criticism is leveled by Machery (2016: 119), which calls the IAT's predictive validity "extremely low". Oswald, Mitchell, Blanton, Jaccard and Tetlock (2015: 562) argues similarly that the mean correlation between implicit measures and behaviors involving race or ethnicity is "small (or very small)". Further, Oswald et al. (2015: 562) faults the claim that implicit biases are associated with significant, real-world consequences for being based on "arguably . . . untenable assumptions".

This article offers rebuttals to this second set of criticisms. I argue the correlations and accompanying effect sizes estimated in implicit bias research suggest implicit racial biases have significant discriminatory effects in at least some circumstances.[1] Section 2 reviews several competing meta-analyses of studies on implicit measures and behaviors and adjudicates, in ways favorable to defenders of implicit bias research, the central methodological and interpretive differences among their authors. In addition, Section 2 advocates a general framework for assigning meaning to the correlations and effect sizes estimated by meta-analyses on implicit bias research. With the aid of this framework, Section 3 offers an in-depth interpretation of the most recent and extensive meta-analysis of studies on implicit measures and intergroup behaviors (Kurdi et al. 2018), including those involving race. According to the interpretation advocated in Section 3, it is plausible that implicit racial biases have significant discriminatory effects at both the individual and group levels.[2] This interpretation is compatible with, but does not entail, the conclusion that implicit racial biases in fact cause such effects.[3]

1. "Significant" is here intended in both a descriptive and normative sense, viz.: (i) as characterizing an effect size above an appropriate triviality threshold for this area of human behavioral research; and (ii) as characterizing behavioral effects that are socially, politically or morally consequential, and, therefore, arguably worth allocating resources toward mitigating.

2. While this article focuses specifically on implicit racial biases and racial discrimination, it may be that its morals generalize to other kinds of implicit attitudes and discrimination, for example those involving gender, sexual orientation or age. Supporting such generalizations, however, is beyond the scope of this article, so I do not assume in what follows that such generalizations hold.

3. An important criticism of implicit bias research, which I cannot adequately address here, is based on an interpretation of Forscher et al.'s (2019) recent meta-analysis of studies of interventions on implicit attitudes (where an "intervention" is any procedure that may change one or more measures of implicit bias). According to this interpretation, implicit attitudes are correlated with,

In Section 4, I consider the role implicit bias research might play in understanding racial inequality on both a non-causal and causal construal. Section 4 does not argue that research on implicit racial bias can provide insight into racial discrimination and inequality equal to or greater than accounts that focus on explicit racial biases or structural causes, such as political and geographic segregation. Rather, according to the defense of implicit bias research offered here, it is plausible that research on implicit racial bias will contribute, likely alongside a variety of research programs, to our best accounts of racial inequity.

1.1. Implicit Associations

Implicit associations are sometimes portrayed as among the most significant discoveries by social psychologists over the past several decades. For our purposes, "associations" may be understood as traces of past experiences that encode evaluations of an object, including social objects. Such associations are often regarded as "biases" in that they encode favorable or unfavorable evaluations of social groups, including social groups defined in terms of race. There is no consensus on how to precisely characterize the sense(s) in which implicit biases are "implicit". However, proposals include: being relatively inaccessible to conscious awareness or introspection (Greenwald, McGhee, & Schwartz 1998), being relatively difficult to bring under reflective control, and contributing to quick and efficient cognitive processing that is more or less "automatic" (Gawronski & De Houwer 2014). Similarly, there is no consensus on the ontology of implicit biases—whether they are a type of belief, attitude, feeling, process, motivation, psychological trait, or situational feature (see Brownstein, Madva & Gawronski 2019 and Holroyd, Scaife, & Stafford 2017 for discussion). Little in this article depends on the resolution of such disagreements over the nature of implicit bias,

but do not cause, behaviors. A "merely correlational" interpretation of Forscher et al.'s findings is both widespread (embraced by Singal 2017; Bartlett 2017; Buckwalter 2018) and *prima facie* plausible in light of Forscher et al.'s findings of no evidence that changes in implicit measures correspond to or mediate changes in behaviors (2019: 540). However, I offer two brief critical comments. First, in their most recent revision (August 2019), Forscher et al. strike multiple cautious notes with respect to a merely correlational interpretation. For example, they maintain their findings do not admit of a "single interpretation" (2019: 544) and that the currently available evidence "cannot decide" between a merely correlational interpretation and one framed in terms of limitations on their methodology or data (2019: 544). (For one potential methodological limitation, see Section 2.3 on meta-analytic inclusion criteria for behavioral effects.) Second, as I argue in Section 4, even supposing implicit racial biases are correlated but not causally related to racially biased behaviors, research on implicit racial bias may nevertheless play a substantive role in accounts of racial discrimination and inequality.

so I am noncommittal in what follows.[4] This article places greater importance on whether the targets of implicit measures, whatever precisely they measure, plausibly have significant behavioral effects.

1.2. Measuring Implicit Bias

A variety of indirect tests have been developed to measure implicit bias, many of which involve timed sorting tasks between visual images or words on the one hand, and positive and negative attributes on the other. The best-known of these is the Implicit Association Test (IAT) (Greenwald et al. 1998), but other measures include The Affective Lexical Priming Score, The Go/No-Go Association Task, The Affect Misattribution Procedure, The Sorting Paired Feature Task, and the Multi-Category Implicit Association Test (see Bar-Anan & Nosek 2014 for discussion). A guiding idea behind many of these indirect tests for implicit associations to social groups is that greater cognitive effort and time is required to match a positive attribute to a social group, or to a member of a social group, when subjects implicitly associate negative attributes with that group, and *vice versa*. Thus, for several of the above, greater response times and error rates on sorting tasks involving positive and negative attributes are interpreted to measure the strength of implicit associations that may encode biases.

Indirect tests of implicit attitudes contrast with direct tests of subjects' explicit attitudes, which typically rely on their self-reports about the relevant attitude. The results of the two testing methods often do not coincide—for example, Hofmann, Gawronski, Gschwendner, Le, and Schmitt (2005: 1376) estimate an average correlation of r = .24 between explicit and implicit measures, with correlations varying significantly across contexts. Based in part on the relatively low correlation between direct and indirect testing results, many social psychologists differentiate implicit from explicit cognition. However, there is no consensus on the extent to which implicit and explicit cognition are distinct or stable features of the human mind,[5] and I will remain neutral on this issue in what follows. The central questions about implicit racial bias this article investigates may be formulated in terms of implicit *measures*, viz.: What predictive relations hold between implicit measures of racial bias and racially biased behaviors; and, are the behavioral effects associated with measures of implicit racial bias significant?

4. The reader is cautioned against inferring any substantive ontological commitments from terminological choices in this article for referring to implicit associations.

5. For example, Schimmack (2019) argues that the low correlation between indirect and direct testing results is better explained by measurement error than by positing distinct types of cognition.

One strategy for answering both questions is by consulting meta-analyses that synthesize the findings of large numbers of studies. This strategy has been widely adopted in discussions on the significance of implicit bias research even though some philosophers of science have discounted the general evidential value of meta-analytic studies. For example, some philosophers have argued meta-analyses merely aggregate potentially biased results (Romero 2016) or merely highlight the analysts' subjective and arbitrary decisions (Stegenga 2011). It is beyond the scope of this article to rebut these general critiques, and with minor exceptions I do not engage them. Instead, this article's investigations proceed on the assumption that meta-analyses are among our best sources of evidence for stochastic effects on large populations (Holman 2019; Bruner & Holman 2019).

2. Meta-Analyses of Studies on Implicit Measures and Behaviors

In this section, I offer a brief historical narrative on meta-analyses over the past decade of studies on implicit measures and behaviors. The narrative is offered to highlight the dynamic nature of implicit bias research and provide perspective on the degree of certainty it is appropriate to assign to meta-analytic findings in this domain. In addition, tracing the methodological and interpretive controversies among authors of competing meta-analyses on implicit bias research is intended to shed light on the potential significance of their findings for real-world settings.[6]

2.1. A Series of Competing Meta-Analyses

The first meta-analysis of studies on implicit measures and behaviors, Greenwald, Poehlman, Uhlmann, and Banaji (2009), encompassed many behavioral domains, including consumer choices, voting, and "intergroup" behaviors, that is, behaviors toward (members of) social groups. According to this meta-analysis, the most common measure of implicit bias—the IAT—could predict, in some contexts better than explicit measures, a variety of behaviors, especially socially sensitive and spontaneous intergroup behaviors involving race and ethnicity (Greenwald et al. 2009: 28). The mean correlations between IAT scores and racial and other intergroup behaviors was, according to Greenwald et al.'s estimate,

6. See Machery and Doris (2017) for advice on appropriately interpreting meta-analyses in psychology.

r = .24 and r = .20 respectively (2009: 28).[7] A second meta-analysis, Cameron, Brown-Iannuzzi, and Payne (2012), estimated an average correlation of r = .28 between sequential priming measures of implicit social cognition (rather than the IAT) and social behaviors. Both meta-analyses supported the validity of implicit measures as predictive tools and potentially implicated implicit racial bias in discriminatory behaviors in a variety of social contexts, including hiring, admissions, healthcare, law enforcement, and other daily interactions.

Subsequent meta-analyses, however, appeared to weaken those conclusions. Instead of correlations in the range of r = .20 to .28, Oswald, Mitchell, Blanton, Jaccard, and Tetlock (2013: 178) estimated a smaller mean correlation of r = .14 between IAT scores and behaviors toward individuals or groups defined in terms of race or ethnicity. This estimate was derived by employing advances in statistical techniques[8] and adopting different inclusion criteria for measured behavioral effects (see Section 2.3 for discussion of these criteria). According to this meta-analysis, the IAT did not add any predictive power to explicit measures for predicting behavior, and in particular was not usefully predictive of spontaneous behaviors (2013: 183). Oswald et al. concluded, "the IAT provides little insight into who will discriminate against whom, and provides no more insight than explicit measures of bias" (2013: 188).

These substantially weaker results sparked controversy over the statistical and societal significance of correlations linking implicit measures to biased behaviors. For example, Machery (2016), drawing on Oswald et al.'s (2013) meta-analysis, urged researchers and others to significantly pare back their estimation of implicit bias's role in explaining discrimination and inequality:

> one may get the erroneous impression that indirect measures are excellent predictors of biased behaviors, since implicit attitudes are called upon to explain many social ills. The truth, however, is quite different . . . since the predictive validity of the implicit association test is extremely low. (Machery 2016: 119)

Research that immediately followed Machery's appraisal appeared to buttress it. For example, Carlsson and Agerström (2016) re-analyzed the same set of studies as Oswald et al. (2013) but operationalized behavioral effects more strictly so as

7. Correlations mentioned in this article should be assumed to be zero-order correlations unless specified otherwise.

8. Specifically, Oswald et al. (2013) abandoned the "summary statistical method" used in Greenwald et al. (2009), opting for a statistical method that seeks to establish probabilistically independent samples where appropriate, while modeling statistical dependencies among effects for samples within the same study (see Oswald et al. 2015 for discussion).

to include only genuinely discriminatory behavioral effects.[9] They reported a similar estimate to Oswald et al. (2013) of the mean correlation between implicit measures and behavioral effects, viz. $r = .15$. More strikingly, Forscher et al. reported "the overall correlation between implicit measures and behavior in our meta-analysis was small. . . ($r = .09$)" (2019: 41).[10] In concert with this trend of diminishing correlations and effect sizes, more strident criticisms of implicit bias research proliferated (e.g., Mitchell & Tetlock 2017; Singal 2017; Bartlett 2017; Mitchell 2018).

2.2. Assessing Competing Meta-Analyses

Authors of competing meta-analyses of studies on implicit measures and behaviors have been divided over two central controversies—one methodological, the other interpretive (cf. Greenwald et al. 2015 and Oswald et al. 2015 for discussion). The methodological controversy is over whether to include, in meta-analytic estimates of a mean correlation between implicit measures and behaviors, behavioral effects that a study's authors did not predict would be correlated with implicit measures. The interpretive controversy is over whether the various mean correlations estimated by competing meta-analyses, regardless of their methodology, suggest implicit biases have significant discriminatory effects. In what follows, I argue the methodological controversy is inconclusive in the sense that both methodological choices—whether to include or exclude behavioral effects based on author prediction—are equally warranted given the available evidence. With respect to the interpretive controversy, I maintain defenders of implicit bias research have substantially stronger arguments for the conclusion that the mean correlations estimated by competing meta-analyses suggest implicit biases have significant discriminatory effects.

9. Carlsson & Agerström's operationalization of discriminatory behavioral effects might be challenged since it requires that race be "isolated from other factors" (2016: 280). For example, they maintain that while the race IAT predicted voting intentions for John McCain and Barack Obama, such voting patterns are not discriminatory because "there are more differences between Obama and McCain than their races" (2016: 280). While this strict operationalization has benefits, it also risks excluding behaviors that have both discriminatory and non-discriminatory effects. An alternate operationalization might classify behaviors as discriminatory if at least one consequence of the behavior disadvantages one or more members of a race compared to another.

10. The first version of Forscher et. al.'s meta-analysis appeared online in 2016. This meta-analysis primarily focuses on studies of interventions of implicit attitudes (see Footnote 4), but also estimates a mean correlation between implicit measures and behaviors. Unlike Oswald et al. (2013), Forscher et. al.'s (2019) scope is not restricted to behaviors involving race or ethnicity, but encompasses intergroup behaviors generally as well as non-social behaviors.

2.3. The Methodological Controversy: Meta-Analytic Inclusion Criteria for Behavioral Effects

Greenwald et al. (2009) and Cameron et al. (2012) (as well as Kurdi et al. 2018; see Section 3) adopt inclusion criteria for behavioral effects that are sensitive to a study's stated predictions; in contrast, Oswald et al. (2013), Carlsson and Agerström (2016), and Forscher et al. (2019) adopt inclusion criteria for behavioral effects that are insensitive to prediction. In other words, the latter but not the former group of meta-analyses include behavioral effects that the studies' authors did not predict would be correlated to measures of implicit associations. According to Greenwald et al. (2015: 556), this methodological difference accounts for a "substantial portion" of the difference between the estimated mean correlation between implicit measures and behavioral effects by Oswald et al. (2013) (r = .14) and Greenwald et al. (2009) (r = .22).[11]

Greenwald et al. (2015) defends Greenwald et al. (2009)'s (and Kurdi et al. 2018's)[12] inclusion criteria on the grounds that authors' expected findings of behavioral effects were often based on such plausible and mundane background hypotheses as:

> (a) measures of attitude toward a group should predict behavior favorable or unfavorable to the group and (b) measures of a stereotype of the group should predict stereotype-consistent judgments or behavior toward members of that group (2015: 554).

On the assumption that such background hypotheses systematically informed prediction of behavioral effects in the studies analyzed, one might reasonably maintain that including measures of behavioral effects that were not predicted to occur—including some studies in which behavioral effects were correctly predicted to be absent—artificially reduces estimates of mean correlations between implicit measures and behaviors. Thus, there are potential benefits associated with the more restrictive inclusion criteria for behavioral effects based on author prediction adopted by Greenwald et al. (2009) (and Kurdi et al. 2018).

11. In particular, Greenwald et al. (2015) estimate these different inclusion criteria account for a little over half the difference (i.e., greater than r = .04) between the two meta-analyses. See table 1 in Greenwald et al. (2015: 555) for a brief presentation of the evidence for this estimate.

12. Kurdi et al. represent (2018: 6) that their inclusion criteria for behavioral effects, which were sensitive to author prediction, were mostly based on the same plausible and mundane background hypotheses highlighted by Greenwald et al. (2015). See Brownstein et al. (2019) for a critique of inclusion criteria for behavioral effects that are insensitive to author prediction, as well as discussion of some distinctive features of Cameron et al. (2012)'s inclusion criteria for behavioral effects that I leave aside here.

But this methodological choice also has potential costs. For example, Oswald et al. (2015) worries that excluding measures of behavioral effects based on researchers' predictions can lead to significant information loss, particularly in light of inconsistent theoretical perspectives across studies that guide such prediction:

> If the meta-analyst . . . defers to the judgment of different researchers in different research reports, inconsistency and the omission of substantial amounts of information can result. This possibility was a specific concern for us, because researchers conducting different IAT studies sometimes embraced different theories (2015: 563).

And indeed, predictions in the studies analyzed by both Oswald et al. (2013) and Greenwald et al. (2009) were guided by hypotheses based on single-association and double-dissociation models, on expectations based on the social sensitivity of the attitude measured, as well as on the spontaneity of the relevant behavior. Given this heterogeneity of theoretical perspectives, one might reasonably adopt a uniform policy on the inclusion of behavioral effects that is insensitive to author prediction, which, as Greenwald et al. (2015: 557) allows, "fits with . . . well-known methodological strategy" in psychological research. Moreover, such inclusion criteria might, arguably, render meta-analytic findings less susceptible to the analysts' bias and less likely to amplify such biases as p-hacking and HARKing[13] by the authors of the studies analyzed. On the basis of such potential strengths, Oswald et al. call their methodological choice less "problematic" (2015: 563) than the alternative inclusion criteria adopted by Greenwald et al. (2009) (and Kurdi et al. 2018).

By contrast, in light of the potential benefits and costs associated with both inclusion criteria, Greenwald et al. (2015: 556) judge that "both strategies were justifiable". Given the evidence and arguments presented by critics and defenders of implicit bias research, this tolerant judgment is surely correct. Both methodologies have potential benefits, and neither has costs that demonstrably outweigh the benefits. Perhaps future inquiry will provide a basis for preferring either methodology. For example, future analysis might quantify the extent to which prediction in the relevant studies of implicit measures and behaviors were guided by plausible and relatively mundane background assumptions as opposed to inconsistent theoretical speculation. But even if such an analysis were performed, it might well be that no clear case would emerge for regarding either

13. "p-hacking" refers to biased data analysis whereby the selective collection or reporting of relationships among variables makes insignificant results appear significant. "HARKing" in contrast refers to hypothesizing by researchers after their results are already known but presenting these hypotheses as formulated prior to any results and subsequently confirmed by them.

strategy as superior. Methodological controversies can be difficult to resolve, even in principle. Distinct methodologies can simply yield different kinds of information that are useful in different contexts and for different purposes.

The inconclusive nature of this methodological controversy might appear to support Stegenga (2011)'s view that competing meta-analyses typically highlight different, arbitrary choices by the analysts. But I believe we may draw a more informative conclusion by scrutinizing the methodologies of competing meta-analyses of studies on implicit measures and behaviors. In particular, it is informative that multiple meta-analyses of studies on implicit measures and behaviors with different but equally warranted methodologies all find correlations and accompanying effect sizes that suggest implicit biases have significant discriminatory effects in at least some circumstances. I argue for this conclusion in the next two subsections.

2.4. The Interpretive Controversy: Interpreting Correlations and Effect Sizes in Implicit Bias Research

Based on Oswald et al.'s (2013) finding of an average correlation of r = .14 between implicit measures and behaviors involving race and ethnicity, Machery (2016) characterizes the IAT's predictive validity for behavioral effects as "extremely low" and Oswald et al. (2015: 565) calls this mean correlation "small by conventional standards" and "small (or very small)" (2015: 562). By contrast, Greenwald et al. (2015: 553) maintains this same "estimated aggregate correlational effect size. . . [is] large enough to explain discriminatory impacts that are societally significant". This raises a question about how to interpret the magnitudes of correlation coefficients and accompanying effect sizes estimated by competing meta-analyses of studies on implicit measures and behaviors.

One strategy for answering is by comparison to other findings. Some comparisons appear favorable to defenders of implicit bias research, since even the lower end of meta-analytic estimates of mean correlations between implicit measures and behaviors are of similar magnitude to other findings in psychology that are considered significant, and consistent with causal relationships regarded as important in other domains. For example, ibuprofen's mean correlation to reducing headache pain has been estimated as r = .14 (Meyer et al. 2001: 131). Similarly-sized correlations accompany a variety of psychological findings also regarded as well-established and significant. According to Richard, Bond, and Stokes-Zoota's (2003) calculations, the mean correlation for the finding that people attribute failures to their bad luck is $r = .10$, that scarcity raises the perceived value of a commodity is $r = .12$, and that informational sources viewed as more credible are more persuasive is $r = .10$ (2003: 354–55). In light of these

comparisons (see also Brownstein, Madva, & Gawronski 2020 for discussion), we should reject the assumption that a mean correlation of r = .14 between implicit measures and behaviors necessarily indicates trivial effects. Such an assumption is inconsistent with a variety of results regarded as significant in psychology and other scientific domains.

On the other hand, alternate comparisons are less favorable to defenders of implicit bias research. For example, the mean correlations between measures of general mental ability and job performance range from $r = .28$ to $r = .38$, depending on how much training the job requires (Schmidt & Hunter 2004: 165). Similarly, the mean correlation for the finding that people are more likely to be aggressive when provoked is $r = .36$ (Richard et al. 2003: 354) and the correlation between IQ scores for young adults (aged 19–23) and occupation type is r = .37 (Strenze 2007: 413). In comparison to these findings, measures of implicit bias fall short. Which comparisons are in fact appropriate benchmarks for judging the correlations and accompanying effect sizes in implicit bias research? By itself, invoking small numbers of comparisons is insufficient to resolve the interpretive controversy. Instead, we might turn to selecting an appropriate, general framework for interpreting the magnitude of correlations and effect sizes in psychological research, including those reported by competing meta-analyses on studies of implicit measures and behaviors.

A standard interpretational framework, widely adopted across the social sciences, is based on guidelines provided by the influential statistician and behavioral science researcher Jacob Cohen (Cohen 1988). Cohen's guidelines are intended to help assign meaning to correlations and effect sizes in many contexts of human behavioral research, and may appear to favor critics of implicit bias research. According to Cohen's guidelines, correlations of r = .5 and above should be regarded as "large"; correlations between r = .3 and r = .5 should be regarded as "medium"; and correlations from r = .1 to r = .3 should be interpreted as "small," with correlations below r = .1 shading into insignificance. Oswald et al.'s (2015: 565) reference to a correlation of r = .14 as "small by conventional standards" likely reflects Cohen's guidelines or a similar framework. These conventional standards imply that the mean correlations between implicit measures and behaviors, as estimated by Oswald et al. (2013), Carlsson and Agerström (2016) and Forscher et al. (2019), were nearing or had crossed the threshold into triviality.

Cohen's guidelines, however, were intended only as rules of thumb whose appropriate application varies by context. In presenting them, Cohen noted the "risk inherent in offering conventional operational definitions . . . in as diverse a field of inquiry as behavioral science" (1988: 25). He further cautioned that " 'small,' 'medium,' and 'large' are relative, not only to each other, but to the area of behavioral science or even more particularly to the specific content and

research method being employed in any given investigation" (1988: 25). Arguably, Cohen's guidelines are often less apposite than an alternate framework for interpreting the correlations and effect sizes found in research on implicit racial biases, since implicit racial biases' effects, if they exist, plausibly range over a large number of events, affecting either large populations or single individuals over long time periods.

That correlation coefficients and effect sizes might require nonstandard and counterintuitive interpretations if large numbers of events are involved was forcefully argued in Abelson (1985), which highlighted the apparently small but nevertheless significant correlation between a professional baseball player's skill and success at bat.[14] Abelson pointed out that the correlation between batting skill and success with respect to any single at-bat for a highly skilled professional baseball player (batting average in the low .300s) as compared to a less skilled player (batting average in the low .200s) might appear quite small. In particular, Abelson calculated the correlation to be r = .056, where the corresponding effect size would thus explain less than 1/3rd of 1% of the variance between the batting outcomes of lower versus more highly skilled players (1985: 131). But the apparently small magnitude of the correlation and accompanying effect size, when considered over hundreds or thousands of at-bats, can nevertheless indicate important consequences, potentially accounting for the difference between baseball teams that win championships and teams with losing seasons.

With respect to the potential effects of implicit racial biases, Greenwald et al. (2015) illustrate via a hypothetical that even the small (by conventional standards) mean correlation between implicit measures and behaviors estimated by Oswald et al. (2013) (r = .14) might indicate a substantial impact at the group level. In particular, they raise the possibility that this correlation might implicate implicit racial biases in such significant discriminatory effects as an excess of nearly 10,000 police stops of racial minorities in New York City each year. Their hypothetical is informed by data but also invokes speculative assumptions, so it should not be confused with a descriptively accurate, causal explanation of New York City policing.[15] Nonetheless, Greenwald et al.'s hypothetical makes

14. The relevance of Abelson's work to interpreting the correlations and effect sizes in implicit bias research is argued in Sripida (2017).

15. Greenwald et al.'s hypothetical should not be confused with a descriptively adequate account of New York City policing in terms of the effects of implicit biases for multiple reasons. First, no study has been conducted that measures the implicit racial biases of New York City police officers and correlates them to police stops. The studies analyzed by Greenwald et al. (2015) and Oswald et al. (2015) contained different kinds of behavioral effect and only further empirical inquiry could establish whether the relevant findings generalize to implicit racial biases and New York City policing (though see Hehman, Flake, & Calanchini 2017 on correlations between regional implicit racial bias and racially disproportionate police fatalities). Second, even were such

a valid if narrow statistical point: the correlations and effect sizes at issue are not inherently indicative of trivial impacts. In their hypothetical, the potential impact of an r = .14 correlation between implicit racial biases and racially biased policing was calculated with data on nearly 200,000 police stops per year across 76 precincts, where New York City police "stopped an average of 38.2% . . . more of each precinct's Black population than of its White population" (Greenwald et al. 2015: 558). Greenwald et al. (2015)'s hypothetical illustrates that given large variability, and a large number of events, even the lower end of mean correlations between implicit measures and behaviors estimated by Oswald et al. (2013), Carlsson and Agerström (2016) and Forscher et al. (2019) might indicate significant effects for society.

Greenwald et al.'s (2015) and Abelson's (1985) examples imply that an alternative to Cohen's guidelines for interpreting correlations and effect sizes in implicit bias research may often be appropriate. One such alternative framework for evaluating psychological research more generally is advocated by Funder and Ozer (2019: 11):

> We offer . . . the following New Guidelines: . . . an effect with the size of *r* = .05 is "very small" for the explanation of single events but potentially consequential in the not-very long run, *r* = .10 is still "small" at the level of single events but potentially more ultimately consequential; an effect size of *r* = .20 is "medium" and of some use even in the short run and therefore even more important; and an effect size of *r* = .30 is "large" and potentially powerful in the short and long run.[16]

Funder and Ozer's framework is likely more apposite than Cohen's guidelines, in many circumstances, for interpreting the estimated mean correlations reported by competing meta-analyses of studies on implicit measures and behaviors. Arguably, this is the case for research on implicit racial biases in particular because they are potentially implicated, over relatively long timeframes, in a multitude of interactions among individuals perceived to be of different races.

---

a study conducted, a number of confounding factors, including structural causes and explicit racial biases, might preempt implicit racial biases' influence on biased policing. However, see Section 4 for argument that it is plausible implicit bias research will play some role in our best accounts of racial discrimination and inequality.

16. Gignac and Szodorai (2016: 74) advocates a similar interpretational framework for individual difference research based on their extensive meta-analysis of published results, whereby r = .10, .20, and .30 should be considered, respectively, as "relatively small, typical, and relatively large" correlations. (See also Hemphill's 2003 alternative to Cohen's guidelines.)

2.5. Racially Biased Behaviors and Cumulative Effects

With respect to assessing the effects of implicit biases at the individual as opposed to group level, establishing that "small" correlations may nevertheless indicate significant effects for an individual often requires establishing that the relevant effects would *accumulate* if repeated over that individual's life. That is, the relevance of implicit bias research for a single individual rather than a population is often determined by whether at least some of the behavioral effects found to be correlated with implicit measures build over time rather than dissipate (see Mallon & Kelly 2012 for discussion). Some kinds of effects do not accumulate. For example, the effects of perceptual and other kinds of stimuli can decrease for an individual despite repetition, as counter-processes of habituation facilitate environmental adaptation. Thus, an individual who moves to a large city can become accustomed to traffic noises rather than experience an accumulation of their initially disruptive effects. An important question for assessing research on implicit racial bias is whether we ought to view the effects of racially biased behaviors that have been correlated to implicit measures as more akin to traffic noises or batting outcomes.

Oswald et al. (2015) are skeptical that the racially biased behaviors studied by implicit bias researchers have real-world, cumulative effects. They remark, "[w]e do not doubt that the effects of small negative events can, in principle, accumulate over time with consequential effects," but warn that "whether this happens for the outcomes studied in IAT research" should not be "simply stipulated" (2015: 568). Oswald et al. go further in their skepticism about the claim that "the small effects found in research laboratories translate into consequential real-world effects," maintaining that such a claim "depend[s] crucially on untested and, arguably, untenable assumptions" (2015: 562).

It is reasonable for Oswald et al. to warn against simply stipulating that the behavioral effects studied in implicit bias research likely accumulate in real-world settings. But the assertion that the accumulation of such effects in real-world circumstances is "untenable" expresses an unwarranted level of skepticism. Implicit bias research has studied, including in field rather than lab settings, at least some behaviors whose effects would plausibly accumulate if repeated. For example, a field study by Rooth (2010) found that greater levels of negative implicit bias in Swedish employers against Arab/Muslim men positively correlated to fewer offers of interviews from those employers. While the correlation is "small" or "very small" by Cohen's guidelines or a similar framework ($r = .113$), this type of effect would plausibly accumulate if repeated, with the capacity to alter the careers and life courses of the individuals affected. Oswald et al. (2015) acknowledge Rooth (2010)'s field study and its potential implications but counsel "caution in making broad claims for real-world meaning from this one study" (2015: 567).

Additional research not analyzed by Oswald et al. (2013) or Oswald et al. (2015), however, undermines their skepticism in two ways. First, additional field studies suggest the finding in Rooth (2010) is not a misleading outlier with respect to identifying plausibly real-world cumulative effects associated with implicit racial biases. Second, the most recent and extensive meta-analysis of studies on implicit measures and intergroup behaviors—Kurdi et al. (2018)—in general supports the external validity of implicit bias research, finding no differences among the analyzed studies as a "function of study setting" (2018: 15). On the first point, Cooper, et al. (2012) conducted a field study of doctors' implicit cognitive stereotyping of patient race and found correlations to "visit length [and] speech speed" (2012: 981). That is, according to this study, doctors on average spend less time and speak faster with patients who are members of implicitly less favored races.[17] Similarly, Hagiwaraet al.'s (2014) field study found that greater levels of anti-Black implicit bias among doctors predicted greater physician-patient talk ratios (i.e., the ratio of time physicians talk to the time patients talk) for Black patients (2014: 127). On the assumption that less time and lower-quality communication during doctor visits can lead to inferior health outcomes, both of these studies identify behavioral effects whose consequences, if repeated, would plausibly accumulate over an individual's life. That is, repeatedly experiencing inferior mental or physical health outcomes would plausibly build over time into significant effects for someone who is affected compared to someone who is not. Thus, as with the impact of implicit racial biases at the group level, there is reason to believe the lower end of meta-analytic estimates of mean correlations between implicit measures and behaviors suggest implicit racial biases may have significant impacts at the individual level.

The upshot, with respect to the controversy over interpreting the magnitude of correlations and effect sizes as reported by competing meta-analyses of studies on implicit measures and behaviors, is that defenders of implicit bias research have a more compelling position. Even the smaller mean correlations between implicit measures and behaviors estimated by Oswald et al. (2013), Carlsson and Agerström (2016), and Forscher et al. (2019) are suggestive of potentially substantial discriminatory effects, as Greenwald et al. (2015) argues, "either because

17. One might worry Cooper et al.'s study engages in p-hacking since it gathers data on many behavioral effects that were not correlated to implicit measures and highlights those behavioral effects that were. Cooper et al. anticipate this worry, however, and respond: "Because the study included multiple comparisons, the possibility of statistical type I error exists; however, this is unlikely because analyses were conceptually driven and grounded in previous literature, most of the observed associations are in the expected directions, and findings across related measures are consistent" (2012: 985). While this response may alleviate some worries related to p-hacking, caution is nevertheless warranted in interpreting Cooper et al.'s results. Thanks to an anonymous reviewer for raising this worry.

they can affect many people simultaneously or because they can affect single persons repeatedly" (2015: 553). This claim is not untenable in light of the best available evidence; it is plausible.

Despite the success of the above replies to critics of implicit bias research, the defense explored so far is only partial. Such replies do not establish that implicit measures have incremental predictive validity over explicit measures for predicting behavior. In other words, the above defense of implicit bias research has not shown that implicit measures add any predictive value to extant explicit measures. Nor does the defense address the worry that subsequent meta-analyses of studies on implicit measures and behavioral effects might follow a trend of finding ever smaller mean correlations and accompanying effect sizes.

3. The Most Recent and Extensive Meta-Analysis on Implicit Measures and Behaviors

At least over the near term, however, such worries have abated rather than multiplied. They are not borne out by the most recent and extensive meta-analysis to date, Kurdi et al. (2018), which analyzes a significantly greater number of studies (six to ten times greater) than previous meta-analyses.[18] Kurdi et al. (2018) incorporates some of the statistical improvements employed by Oswald et al. (2013) over Greenwald et al.'s (2009) meta-analysis,[19] but finds that the IAT and other implicit measures have incremental predictive validity for intergroup behaviors, including those involving race, compared to explicit measures (2018: 19). Indeed, Kurdi et al. found that both implicit and explicit measures of intergroup cognition make similarly-sized but unique contributions to predicting intergroup behaviors.[20] In addition, this meta-analysis estimates robust mean correlations under some conditions between implicit measures and intergroup behaviors, with its authors emphasizing that such correlations range as high as r =.37, but with an average correlation across all conditions of r = .10 (2018: 13).

Given the strength of their results, Kurdi et al. tentatively suggest reframing the debate over relations between implicit associations and intergroup behaviors:

> . . . instead of asking *whether* implicit measures of intergroup cognition are related to measures of intergroup behavior, it may be more appropriate

18. The scope of Kurdi et al.'s meta-analysis encompasses intergroup attitudes and behaviors generally, and is thus broader than Oswald et al. (2013), which focused specifically on racial and ethnic attitudes and behaviors, but narrower than Greenwald et al. (2009), which encompassed non-social behaviors such as consumer choices.

19. See Footnote 8 for discussion.

20. See Supplement 6 p. 3 to Kurdi et al. (2018).

> to ask *under what conditions* the two are more or less highly correlated (2018: 7, italics in original).

I will assess this suggested reframing by evaluating the relevance of such conditions to characterizing relations between measures of implicit racial bias and racially biased behaviors.

3.1. Interpreting Kurdi et al. (2018)'s Findings

One might object that Kurdi et al.'s suggested reframing invites implicit bias researchers to cherry-pick a subset of variables associated with higher mean correlations between implicit measures and behaviors.[21] Two points in response are worth considering. First, as I argued above, a mean correlation of r = .10 between implicit measures and intergroup behaviors, which Kurdi et al. find holds across all conditions, plausibly indicates significant effects at both the individual and group levels. Second, there are grounds for adopting Kurdi et al.'s tentative reframing. Among the most compelling is that a similar frame has been widespread for more than half a century in attitude psychology generally, where self-reports of explicit attitudes are widely understood to be better predictors of behaviors under some (theoretically expected) conditions than others (see Brownstein et al. 2020: 5–6, for discussion). In this light, Kurdi et al.'s suggested reframing with respect to research on implicit intergroup attitudes enjoys some plausibility. Ultimately, however, whether it is appropriate to emphasize conditions among Kurdi et al.'s findings that are associated with a higher estimated mean correlation between implicit measures and intergroup behaviors depends on both the nature of the variables that were found to impact Kurdi et al.'s estimate and those that were not.

With respect to the latter, Kurdi et al. report that ten variables they tested were not associated with a higher mean correlation than r = .10.[22] While this might appear to lend credence to the objection that Kurdi et al.'s suggested reframing invites biased selection of variables, it is worth emphasizing that five of the variables found not to be associated with a higher mean correlation between implicit measures and intergroup behaviors would have undermined Kurdi et al.'s findings if they had been. That is, half of the variables that did not impact Kurdi et al.'s estimated mean correlation essentially tested for various forms of bias and invalidity in implicit bias research, including: *publication status* (published

21. Thanks to an anonymous reviewer for raising this objection.

22. See Kurdi et al. (2018: 5–6)'s Supplement 6, available online, for a full report of the variables they tested for potential impact on their estimated mean correlation between implicit measures and intergroup behaviors.

vs. unpublished), *sample population* (general, online, preselected, real-world, or student), *sample origin* (US vs. foreign), *sample composition* (only stigmatized participants, only non-stigmatized participants, or both), and *study location* (lab, online, or real-world) (Kurdi et al. 2018: 5). In the present context, this should reduce (but not eliminate) concerns related to biased selection of variables by Kurdi et al. (2018). Half of the variables that might support such a worry were in fact components of successful tests for flaws in implicit bias research, including publication bias, external invalidity, and various forms of selection bias.[23] Evaluating Kurdi et. al.'s suggested reframing also requires considering the variables that, according their findings, define conditions with a higher mean correlation than r = .10 between implicit measures and intergroup behaviors, including those involving race.

With respect to these moderating variables, the most informative interpretation of Kurdi et al.'s findings, for our purposes, highlights that variables they associate with a mean correlation of r = .23 between implicit measures and intergroup behaviors are plausibly associated with greater measurement accuracy and reflective of many attitudes and behaviors involving race in real-world settings. Additional variables associated with even higher mean correlations may enjoy a similar status, although their status is less clear. On this basis, as well as on the basis of general plausibility considerations on correlations and effect sizes in psychological research, I endorse an interpretation of Kurdi et al.'s findings as identifying one or more informative conditions under which mean correlations between implicit measures and many intergroup behaviors, including those involving race, are near r = .23 but likely below r = .37. Importantly, mean correlations between implicit measures and intergroup behaviors in this range imply significant behavioral effects according to all of the interpretational guidelines we have discussed, including Cohen's.

Kurdi et al. identify a mean correlation of r = .23 between implicit measures and intergroup behaviors for studies satisfying the following two criteria (2018: 23):

(1) behaviors are measured using relative rather than absolute categories
(2) indirect tests for implicit associations use "high polarity" attributes as stimuli

23. Unfortunately, Kurdi et al. did not follow a recent trend of pre-registering predictions about the variables they tested, a trend designed to address such biases as p-hacking and HARKing in published research. According to lead author Benedik Kurdi, work on their meta-analysis began in 2013, essentially before pre-registration was practiced to any significant degree in psychological research (Benedik Kurdi, personal communication). It is to be hoped that pre-registration will improve future analyses of implicit bias research. Thanks to an anonymous reviewer for raising this query concerning pre-registration.

(1) requires that behaviors be measured using more than two categories—for example, donating money, potentially of different amounts, to a White or Black student group as opposed to "absolute" categories, such as a "Yes" or "No" response (2018: 15). Given that the IAT and other implicit measures are themselves relative measures, measuring behaviors using relative rather than absolute categories plausibly allows for more precise calculations of correlations between the two. Further, behaviors appropriately characterized with more than two categories are also plausibly implicated in large numbers of interactions with the potential for discrimination in real-world settings, such as ranking multiple candidates for employment or admissions, as well as determining differently sized allocations of money, healthcare, time or other resources for individuals or groups perceived to be of different races. Thus, (1) is plausibly associated with both greater measurement accuracy and greater ecological validity.

(2) requires that "high polarity" attributes be used as stimuli in indirect tests (e.g., fat vs. thin) rather than attributes that are merely different but less opposed (e.g., sad vs. angry). Kurdi et al. speculate that "high-polarity attributes . . . may produce larger effects than low-polarity attributes . . . because they tap into a more cohesive network of mental representations" (2018: 15). Whatever the reason for a higher mean correlation between implicit measures and intergroup behaviors given (2), indirect tests such as the race IAT often use attributes whose differences exhibit high polarity when measuring negative and positive implicit associations to racial groups. Their results suggest the existence in many individuals of implicit racial attitudes with associated attributes that exhibit high polarity. This in turn supports a correlation of r = .23 (and perhaps higher) as informatively characterizing relations between many implicit racial biases and racially biased behaviors.

Kurdi et al. identify three additional criteria such that, if the studies analyzed satisfy them in addition to (1) and (2), the mean correlation between implicit measures and intergroup behaviors rises to r = .37 (2018: 23). But the relevance of these additional criteria to measurement accuracy or external validity is less clear than for (1) and (2). Kurdi et al.'s first additional criterion requires a standard IAT be used rather than an alternate indirect test. It may be that the IAT is a superior measurement tool to other indirect measures, perhaps because of its greater internal consistency (cf. Bar-Anan & Nosek 2014: 675). On the other hand, it is unclear why variants designed to improve on the IAT would fail to do so. Thus, the import of this criterion is more suggestive than probative.

Kurdi et al.'s second additional criterion requires a "high correspondence" between implicit attitudes and behaviors. This criterion is intended to be sensitive to Ajzen and Fishbein's (1977) principle of correspondence, according to which attitudes are better predictors of behavior when there is clear correspondence between the content of an attitude and the behavior being measured. This

criterion may also be associated with many relations between implicit racial attitudes and behaviors outside an experimental setting. However, Kurdi et al. report blind coding of this variable disagreed significantly with non-blind coding (2018: 21). So again, caution is warranted in assessing the import of this second criterion. A final criterion requires that studies have a "declared focus" on identifying correlations between implicit measures and intergroup behaviors. While this may be of interest as a methodological moderator for a meta-analysis, it does not directly bear on greater measurement accuracy or external validity. Thus, despite a mean correlation of r = .37 featuring in Kurdi et al.'s abstract (and in subsequent references to their meta-analysis, e.g., Brownstein et al. 2019: 9), the grounds are tenuous for regarding a correlation of this magnitude as informatively characterizing relations between measures of implicit associations and intergroup behaviors, including those involving race.

Instead, there is some reason to be skeptical given that a mean correlation of r = .37 between implicit measures and intergroup behaviors might render implicit bias research implausibly successful. As Funder and Ozer comment, in light of the general complexity of relations between cognition and behavior, as well as the history of psychological research:

> A "very large" effect size (*r* = .40 or greater) in the context of psychological research is, we suggest, likely to be a gross overestimate that will rarely be found in a large sample or in a replication. Smaller effect sizes are not merely worth taking seriously. They are also more believable. (2019: 11)

In this light, mean correlations between implicit measures and intergroup behaviors near r = .23, and perhaps ranging modestly higher, are more plausible than a mean correlation of r = .37.[24] Nevertheless, according to the interpretation of Kurdi et al. (2018) I advocate, this meta-analysis has identified one of the highest ranges to date of mean correlations between implicit measures and intergroup behaviors, including those involving race. Moreover, according to reasonable interpretational frameworks advocated by Funder and Ozer (2019) and Gignac and Szodorai (2016), these correlations and accompanying effect sizes are in the "medium" to "large" range for human behavioral research, and suggest implicit social biases, including implicit racial biases, correspond to significant behavioral effects at both the individual and group levels.

24. This range also falls squarely within average correlations found more generally in individual differences research. According to Gignac and Szodorai's (2016) extensive meta-analytic review of the personality and social psychology literature, the average published correlation is r = .19, where r = .11 falls at the 25th percentile and r = .29 at the 75th. Thus, the interpretation of Kurdi et al.'s (2018) findings I advocate places studies on implicit bias solidly within their domain of research, broadly construed.

Kurdi et al.'s (2018) findings support implicit bias research in other important ways, including defending it from worries related to publication bias and external validity. More specifically, Kurdi et al. (2018: 10) found that the unpublished materials they reviewed generally showed slightly larger effects than published results. They also tested for greater model fit under the assumption that unpublished results showed smaller effect sizes than published results, yet found no such fit (2018: 10).[25] On these bases, Kurdi et al. conclude: "unlike other fields of psychology, the study of implicit–criterion relationships is unlikely to be plagued by a widespread file drawer problem" (2018: 10). In this way, Kurdi et al. (2018) at least partially defend their findings from the general worry that meta-analyses often merely aggregate the results of biased studies (Romero 2016). With respect to external validity, Kurdi et al. analyzed whether correlations between implicit measures and behaviors were affected when tested for "under the less controlled conditions of online and field studies" (2018: 15). And their meta-analysis found "no difference" in average correlations between implicit measures and intergroup behaviors "as a function of study settings" (2018: 15). Thus, as with publication bias, their findings defend implicit bias research from the worry that correlations between implicit measures and intergroup behaviors, including those involving race and ethnicity, hold solely in experimental as opposed to real-world settings. In the next subsection, I situate these findings within the context of competing meta-analyses of implicit bias research and the debate over whether implicit racial biases likely have significant discriminatory effects.

3.2. Taking Stock of Competing Meta-Analyses of Studies on Implicit Measures and Behaviors

Let us take stock of the trajectory of meta-analyses of studies on implicit measures and behaviors over the past decade. Despite a mid-decade trend in which meta-analytic estimates of mean correlations between implicit measures and behaviors were diminishing, in fact meta-analyses during this time did not report mean correlations or accompanying effect sizes that would be insignificant for large populations in the short term or for individuals over longer time

25. One might worry that high levels of heterogeneity among the effects in the studies Kurdi et al. analyzed render their tests for publication bias unreliable (thanks to an anonymous reviewer for raising this worry). However, at least one test Kurdi et al. performed for detecting publication bias is plausibly reliable given the degree of heterogeneity they estimate. In particular, Kurdi et al. estimate heterogeneity of $\tau = .14$ in the studies they analyzed, where the test for publication bias they performed that is developed in Vevea and Hedges (1995) is plausibly reliable for levels of heterogeneity up to $\tau = .20$ (personal communication, Benedik Kurdi; Supplement 6, p. 2). This test, as well as four other tests for publication bias Kurdi et al. also conducted, each with different strengths and weaknesses, converged on an identical result of no publication bias.

frames. Now, the most extensive meta-analysis to date has estimated greater mean correlations under conditions plausibly reflective of many real-life circumstances, and where measurement of implicit biases and intergroup behaviors, including those involving race, may have higher degrees of accuracy. Further, this meta-analysis has portrayed implicit measures as adding significant predictive power to explicit measures and defended research on implicit bias from criticisms relating to publication bias and external validity.

While the positive view of implicit bias research just sketched is warranted in light of the best available evidence, a cautious stance is nevertheless appropriate given that research on implicit social cognition is an active, unsettled domain of inquiry. Kurdi et al. themselves highlight systematic inadequacies in implicit bias research, including inattention to measurement error and problematic levels of internal consistency (2018: 16). Moreover, their meta-analysis relies on methodological inclusion criteria for behavioral effects that are sensitive to author prediction. These inclusion criteria are reasonable, but not necessarily superior to alternatives associated with smaller estimated mean correlations between implicit measures and behaviors.

Nevertheless, Kurdi et al. (2018) provides an important defense of research linking implicit measures to intergroup behaviors, including those involving race. While their meta-analysis should not be taken as the last word on correlations between the two, nor on the uniquely predictive powers of implicit measures of bias, neither should its substantial evidential force be dismissed. Kurdi et al.'s findings clearly suggest that measures of implicit racial bias significantly correspond to racially biased behaviors. In broad terms, the best available evidence from social psychology supports the conclusion that implicit racial biases have discriminatory effects that are significant at both the individual and group levels. However, the research on which this conclusion is based primarily investigates questions of correlation rather than causation. Consequentially, I will consider the role research on implicit racial bias might play in accounts of racial discrimination and inequality both on the assumption that implicit racial biases causally influence racially biased behaviors and on the assumption they do not.

4. Implicit Bias Research and Accounts of Racial Discrimination and Inequality

4.1. A Causal Interpretation of Research on Implicit Racial Bias

According to a "causal interpretation" of research on implicit racial bias, findings on correlations between implicit racial biases and racially biased behaviors track, in some cases, the causal influences of the former. For example, a causal

interpretation of Kurdi et al.'s (2018) finding that implicit measures of bias have incremental predictive validity compared to explicit measures implies that, in some cases, implicit biases make unique causal contributions to biased behaviors. Thus, on a causal construal, implicit bias research suggests implicit racial biases are significant causal antecedents of at least some racially biased behaviors. Plausibly, understanding the causes and effects of racially biased behaviors can further our understanding racial inequality. In which case, research on implicit racial bias, on a causal interpretation, provides reason to anticipate that it will contribute to such an understanding.

In addition, a causal construal of implicit racial biases suggests the study of their manipulation and mitigation may play a useful role in efforts to reduce discrimination and ameliorate inequality. I am not here arguing that research on implicit racial bias could provide insight into racial discrimination and inequality equal to or greater than alternative accounts, such as those focused on explicit racial biases or non-psychological, structural causes. Rather, I maintain it is plausible that research on implicit racial biases, if a causal interpretation is apposite, will play a role, likely alongside other kinds of research, in our best accounts of racial inequity and efforts to address it.

4.2. A Non-Causal Interpretation of Research on Implicit Racial Bias

On a non-causal interpretation of implicit racial bias research, implicit racial biases are significantly correlated but not causally related to racially biased behaviors, perhaps because both are products of common social-environmental causes. Compared to a causal interpretation of implicit bias research, a non-causal interpretation casts implicit bias research in a less important light. For example, if implicit racial biases are construed non-causally, research on their short-term manipulation (Lai et al. 2014; Lai et al. 2016) would likely be of limited relevance to accounts of racial inequality or efforts to address it. Nonetheless, I believe we may expect, even on a non-causal interpretation, that implicit bias research will play a useful if limited contributory role.

A central reason for such an expectation is that variables that are correlated but not causally related to significant effects are nevertheless often used to predict them. As Forscher et al. point out, in their meta-analysis that many interpret as supporting a non-causal construal of implicit attitudes, demographic variables such as life expectancy are often used "to predict consequential outcomes within a population, despite lacking causal force themselves" (2019: 544). Forscher et al. anticipate in particular that even if implicit biases are "causally inert . . . implicit measures could be used to predict the prevalence of . . . behaviors within a population" (2019: 544). Such a use for implicit measures is plausible in light of Kurdi

et al. (2018)'s finding that implicit measures have, in general, incremental predictive validity compared to explicit measures for predicting intergroup behaviors. As applied to implicit measures of racial bias, their finding implies that these measures, in some circumstances, better predict racially biased behaviors than explicit measures.

Hehman et al.'s (2017: 395) findings illustrate predictive power of this kind. Their study found that implicit but not explicit measures of White prejudice and stereotype in a region predict disproportional use of lethal force by police on Blacks in that region, relative to regional base rates of Blacks in the population. On a non-causal construal of implicit racial biases, which Hehman et al. consider, implicit measures of White racial prejudice and stereotyping in a region may nevertheless register useful information about the actual causes of law enforcement's disproportionate lethal force on Blacks (2017: 398). Plausibly, understanding racial inequality can be furthered by understanding the causes and effects of such racially biased behaviors as law enforcement's racially disproportionate use of lethal force. Thus, there appears little reason to eschew tools with unique powers for predicting such behaviors. Instead, even if implicit racial biases are construed non-causally, the current state of implicit bias research suggests implicit measures of racial bias may well play a substantive role in our best explanatory accounts of racial discrimination and inequality.

5. Conclusion

I have attempted to avoid overstating either the certainty or uncertainty of conclusions that can be drawn from research on implicit racial biases and their relations to racially biased behaviors. However, I have argued that prominent forms of skepticism about research on implicit racial bias are unwarranted in light of its current state. According to the best interpretation of the evidence currently available, I have argued, implicit racial biases are importantly associated with racially biased behaviors whose effects are plausibly significant at both group and individual levels. This evidence in turn suggests research on implicit racial bias will play a substantive role, likely alongside a variety of distinct kinds of research, in our best efforts to understand, and perhaps ameliorate, racial inequity.

Acknowledgements

For insightful comments on previous drafts of this article, I would like to thank Bennett Holman, Carl Voss, Michael Michael, David Fuller, Sharon Fuller, and

two anonymous referees for this journal. In addition, Benedik Kurdi's input advanced this project considerably.

References

Abelson, R. (1985). A Variance Explanation Paradox: When a Little Is a Lot. *Psychological Bulletin, 97*(1), 129–33. https://doi.org/10.1037/0033-2909.97.1.129

Ajzen, I. and M. Fishbein (1977). Attitude-Behavior Relations: A Theoretical Analysis and Review of Empirical Research. *Psychological Bulletin, 84*, 888–918.

Anderson, E. (2012). Epistemic Justice as a Virtue of Social Institutions. *Social Epistemology*, *26*(2), 163–73. https://doi.org/10.1080/02691728.2011.652211

Bar-Anan, Y. and B. Nosek (2014). A Comparative Investigation of Seven Implicit Measures of Social Cognition. *Behavior Research Methods*, *46*(3), 668–88. https://doi.org/10.3758/s13428-013-0410-6

Bartlett, T. (2017). Can We Really Measure Implicit Bias? Maybe Not. *The Chronicle of Higher Education*. Retrieved from http://www.chronicle.com/article/Can-We-Really-Measure-Implicit/238807

Blanton, H. and J. Jaccard (2006). Arbitrary Metrics in Psychology. *American Psychologist*, *61*, 27–41.

Brownstein, M., M. Madva, and B. Gawronski (2019). What Do Implicit Measures Measure? *WIREs Cognitive Science*, 1–13: *e1501*. https://doi.org/10.1002/wcs.1501

Brownstein, M., M. Madva, and B. Gawronski (2020). Understanding Implicit Bias: Putting the Criticism into Perspective. *Pacific Philosophical Quarterly*, *101*(2), 276–307.

Bruner, J. and B. Holman (2019). Self-Correction in Science: Meta-Analysis, Bias and Social Structure. *Studies in History and Philosophy of Science Part A*, *78*, 93–97.

Buckwalter, W. (2019). Implicit Attitudes and the Ability Argument. *Philosophical Studies, 176*, 2961–90.

Carlsson, R. and J. Agerström (2016). A Closer Look at the Discrimination Outcomes in the IAT Literature. *Scandinavian Journal of Psychology*, *57*, 278–87.

Cameron, C., J. Brown-Iannuzzi, and B. Payne (2012). Sequential Priming Measures of Implicit Social Cognition: A Meta-Analysis of Associations with Behavior and Explicit Attitudes. *Personality and Social Psychology Review*, *16*, 330–50.

Charlesworth, E. and M. Banaji (2019). Patterns of Implicit and Explicit Attitudes: I. Long-Term Change and Stability From 2007 to 2016. *Psychological Science*, *30*(2), 174–92.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Lawrence Erlbaum Associates.

Cooper, L., D. Roter, K. Carson, M. Beach, J. Sabin, A. Greenwald, and T. Inui (2012). The Associations of Clinicians' Implicit Attitudes about Race With Medical Visit Communication and Patient Ratings of Interpersonal Care. *American Journal of Public Health, 102*(5), 979–87.

Forscher, P., C. Lai, J. Axt, C. Ebersole, M. Herman, P. Devine, and B. Nosek (2019). A Meta-Analysis of Procedures to Change Implicit Measures. *Journal of Personality and Social Psychology: Attitudes and Cognition*, *117*(3), 522–59.

Funder, D. and D. Ozer (2019). Evaluating Effect Size in Psychological Research: Sense and Nonsense. *Advances in Methods and Practices in Psychological Science*, *2*(2), 156–68.

Gawronski, B. and J. De Houwer (2014). Implicit Measures in Social and Personality Psychology. In H. T. Reis, C. M. Judd (Eds.), *Handbook of Research Methods in Social and Personality Psychology* (2nd ed., 281–308). Cambridge University Press.

Gignac, G. and E. Szodorai (2016). Effect Size Guidelines for Individual Differences Researchers. *Personality and Individual Differences*, *102*, 74–78.

Greenwald, A. and M. Banaji (1995). Implicit Social Cognition: Attitudes, Self-Esteem, and Stereotypes. *Psychological Review*, *102*(1), 4–27.

Greenwald, A., D. McGhee, and J. Schwartz (1998). Measuring Individual Differences in Implicit Cognition: The Implicit Association Test. *Journal of Personality and Social Psychology*, *74*, 1464–80.

Greenwald, A., T. Poehlman, E. Uhlmann, and M. Banaji (2009). Understanding and Using the Implicit Association Test: III. Meta-Analysis of Predictive Validity. *Journal of Personality and Social Psychology*, *97*, 17–41.

Greenwald, A., M. Banaji, and B. Nosek (2015). Statistically Small Effects of the Implicit Association Test Can Have Societally Large Effects. *Journal of Personality and Social Psychology*, *108*(4), 553–61.

Hagiwara, N., L. Penner, R. Gonzalez, S. Eggly, J. Dovidio, S. Gaertner, . . . T. Albrecht (2014). Racial Attitudes, Physician-Patient Talk Time Ratio, and Adherence in Racially Discordant Medical Interactions. *Social Science & Medicine*, *87*, 123–31.

Hehman, E., J. Flake, and J. Calanchini (2017). Disproportionate Use of Lethal Force in Policing Is Associated with Regional Racial Biases of Residents. *Social Psychological and Personality Science*, *9*(4), 393–401. https://doi.org/10.1177/1948550617711229

Hemphill, J. (2003). Interpreting the Magnitudes of Correlation Coefficients. *American Psychologist*, *58*(1), 78–80.

Hofmann, W., B. Gawronski, T. Gschwendner, H. Le, and M. Schmitt (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.

Holman, B. (2019). In Defense of Meta-Analysis. *Synthese*, *196*(8), 3189–211. https://doi.org/10.1007/s11229-018-1690-2

Holroyd, J., R. Scaife, and T. Stafford (2017). What is Implicit Bias? *Philosophy Compass*, *12*, 12437. https://doi.org/10.1111/phc3.12437e

Kurdi, B., A. Seitchik, J. Axt, T. Carroll, A. Karapetyan, N. Kaushik, . . . M. Banaji (2018). Relationship Between the Implicit Association Test and Intergroup Behavior: A Meta-Analysis. *American Psychologist, 74*(5), 569–86. https://doi.org/10.1037/amp0000364

Lai, C., M. Marini, S. Le, C. Cerruti, J. Shin, J. Joy-Gaba, . . . B. Nosek (2014). Reducing Implicit Racial Preferences: I. A Comparative Investigation of 17 Interventions. *Journal of Experimental Psychology: General*, *143*(4), 1765–85.

Lai, C., A. Skinner, E. Cooley, S. Murrar, M. Brauer, T. Devos, . . . B. Nosek (2016). Reducing Implicit Racial Preferences: II. Intervention Effectiveness across Time. *Journal of Experimental Psychology: General*, *145*(8), 1001–16.

MacDonald, H. (2017). The False 'Science' of Implicit Bias. *Wall Street Journal*. https://www.wsj.com/articles/the-false-science-of-implicit-bias-1507590908

Machery E. (2016) De-Freuding Implicit Attitudes. In M. Brownstein and J. Saul (Eds.), *Implicit Bias and Philosophy: Volume 1, Metaphysics and Epistemology* (104–29). Oxford University Press.

Machery, E. (2017). Should We Throw the IAT on the Scrap Heap of Indirect Measures? Comment on The Brains Blog, January 17. http://philosophyofbrains.com/2017/01/17/how-can-we-measure-implicit-bias-a-brains-blog-roundtable.aspx

Machery, E. and J. M. Doris (2017). An Open Letter to Our Students: Doing Interdisciplinary Moral Psychology. In B. G. Voyer and T. Tarantola (Eds), *Moral Psychology: A Multidisciplinary Guide* (119–43). Springer.

Mallon, M. and D. Kelly (2012). Making Race Out of Nothing: Psychologically Constrained Social Roles. In Harold Kincaid (Ed.), *The Oxford Handbook of Philosophy of Social Science* (507–29). Oxford University Press.

Meyer, G., S. Finn, L. Eyde, G. Kay, K. Moreland, R. Dies . . . G. Reed (2001). Psychological Testing and Psychological Assessment: A Review of Evidence and Issues. *American Psychologist, 56*(2), 128–65. https://doi.org/10.1037/0003-066X.56.2.128

Mitchell, G. and P. Tetlock (2017). Popularity as a Poor Proxy for Utility: The Case of Implicit Prejudice. In S. O. Lilienfeld and I. D. Waldman (Eds.), *Psychological Science under Scrutiny: Recent Challenges and Proposed Solutions* (164–95). Wiley.

Mitchell, G. (2018). An Implicit Bias Primer. *Virginia Journal of Social Policy and the Law, 25*(1), 28–55.

Oswald, F., G. Mitchell, H. Blanton, J. Jaccard, and P. Tetlock (2013). Predicting Ethnic and Racial Discrimination: A Meta-Analysis of IAT Criterion Studies. *Journal of Personality and Social Psychology*, *105*(2), 171–92.

Oswald, F., G. Mitchell, H. Blanton, J. Jaccard, and P. Tetlock (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), 562–71. https://doi.org/10.1037/pspa0000023

Richard, F., C. Bond, and J. Stokes-Zoota (2003). One Hundred Years of Social Psychology Quantitatively Described. *Review of General Psychology*, *7*(4), 331–63.

Romero, F. (2016). Can the Behavioral Sciences Self-Correct? A Social Epistemic Study. *Studies in History and Philosophy of Science Part A*, *60*, 55–69.

Rooth, D. (2010). Automatic Associations and Discrimination in Hiring: Real World Evidence. *Labour Economics, 17,* 523–34.

Schimmack, U. (2019). The Implicit Association Test: A Method in Search of a Construct. *Perspectives on Psychological Science,* 1745691619863798.

Schmidt, F. and J. Hunter (2004). General Mental Ability in the World of Work: Occupational Attainment and Job Performance. *Journal of Personality and Social Psychology*, *86*(1), 162–73.

Singal, J. (2017). Psychology's Favorite Tool for Measuring Racism Isn't Up To the Job. New York Magazine. Retrieved from http://nymag.com/scienceofus/2017/01/psychologys-racism-measuring-tool-isnt-up-to-the-job.html

Sripida, C. (2017). *Putting "Tiny" Correlations Between Implicit Attitudes and Behavior in Perspective* http://philosophyofbrains.com/2017/01/17/how-can-we-measure-implicit-bias-a-brains-blog-roundtable.aspx

Stegenga, J. (2011). Is Meta-Analysis the Platinum Standard of Evidence? *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, *42*(4), 497–507.

Strenze, T. (2007). Intelligence and Socioeconomic Success: A Meta-Analytic Review of Longitudinal Research. *Intelligence*, *35*(5), 401–26.

Vevea, J. and L. Hedges (1995). A General Linear Model for Estimating Effect Size in the Presence of Publication Bias. *Psychometrika*, *60*(3), 419–35. https://doi.org/10.1007/BF02294384