UNDERSTANDING IMPLICIT BIAS: PUTTING THE CRITICISM INTO PERSPECTIVE

BY

MICHAEL BROWNSTEIN, ALEX MADVA,
AND BERTRAM GAWRONSKI

Abstract: What is the status of research on implicit bias? In light of meta-analyses revealing ostensibly low average correlations between implicit measures and behavior, as well as various other psychometric concerns, criticism has become ubiquitous. We argue that while there are significant challenges and ample room for improvement, research on the causes, psychological properties, and behavioral effects of implicit bias continues to deserve a role in the sciences of the mind as well as in efforts to understand, and ultimately combat, discrimination and inequality.

1. *Introduction*

What is the status of research on implicit bias? Criticism is ubiquitous. Recent meta-analytic reviews suggest that the Implicit Association Test (IAT) is a 'poor' predictor of behavior (Oswald *et al*. 2013) and that changes in scores on implicit measures may not be associated with changes in behavior (Forscher *et al*. 2019). Prominent philosophers have questioned the validity of research on implicit social cognition altogether (e.g., Buckwalter 2018; Hermanson 2017a,b, 2018; Machery 2016, 2017a,b; Yao and Reis-Dennis n.d.). Edouard Machery (2017b), for example, describes an ongoing 'rescue mission' within the field, implying that the relevant research is in peril of being discredited. Machery argues that leading methods for studying and theorizing about implicit bias need to be rethought from the ground up, writing

Pacific Philosophical Quarterly 101 (2020) 276–307 DOI: 10.1111/papq.12302

that we should not 'build theoretical castles on such quicksand'.[1] Headlines in the popular press have been even more pointed. *New York Magazine* reports 'Psychology's Favorite Tool for Measuring Racism Isn't Up to the Job' (Singal 2017); the *Chronicle of Higher Education* asks 'Can We Really Measure Implicit Bias? Maybe Not' (Bartlett 2017); and most pointedly, the *Wall Street Journal* describes 'The False "Science" of Implicit Bias' (Mac Donald 2017).

We argue that while there are significant challenges and ample room for improvement, research on the causes, psychological properties, and behavioral effects of implicit bias continues to deserve a role in the sciences of the mind as well as in efforts to understand, and ultimately combat, discrimination and inequality. In what follows, we first describe the central issues that have been described as crises, anomalies, or puzzles for the field (Section 2). To demonstrate that these alleged anomalies are empirical questions on which progress is steadily being made, we place them in the broader historical context of theorizing on the relationship between attitudes and behavior (Sections 3–4). We respond to potential criticism (Section 5) and then, finally, point to directions for future research (Section 6). Along the way, we highlight the importance of these issues for fundamental questions about the architecture of the mind and the metaphysics of action, especially how mental states, attitudes, and dispositions interact with contextual factors to produce behavior. Specifically, we aim to make progress toward a person-by-situation interactionist theory of the mind and action, which requires rethinking the premises underlying enduring philosophical debates about the importance of personal variables (such as beliefs, traits, or even virtues) and situational opportunities and constraints (including social and environmental factors).

A quick note: Our focus is on the psychology of implicit bias and the psychometric properties of implicit measures. Philosophers, legal theorists, activists, and other social scientists have raised a number of important critical questions about research on implicit bias that we do not address directly here. Perhaps the most well known of these is that research on implicit bias obscures the 'structural' causes of inequality and discrimination (e.g., Banks and Ford 2008; Dixon *et al.* 2012; Haslanger 2015). We have addressed some of these issues elsewhere (Madva 2016b, 2017; Brownstein, Forthcoming; Gawronski and Bodenhausen 2017) and will note links between the issues presented here and these broader concerns where possible.

2. *Central criticisms*

Current criticism is rooted in two sets of findings.[2] The first concerns the extent to which implicit measures predict behavior. The second concerns the stability of individuals' scores on implicit measures over time.

2.1. PREDICTING BEHAVIOR

Estimates of average correlations between individuals' scores on implicit measures and measures of behavior have varied, from approximately $r = 0.14$ to $r = 0.28$ (Cameron *et al.* 2012; Greenwald *et al.* 2009a; Oswald *et al.* 2013). This variety is due to a number of factors, including the type of measures, type of attitudes measured (e.g., attitudes in general versus intergroup attitudes in particular), inclusion criteria for meta-analyses, and statistical meta-analytic techniques. We discuss some of the ramifications of these differences below. Nevertheless, according to standard conventions, all of these correlations are considered small to small-to-medium. Kurdi and Banaji (2017) report that these correlations mean that individual differences in implicit bias account for between 1% and 8% of variance in intergroup discrimination. From these data, critics have concluded that implicit measures, in particular, the IAT (Greenwald *et al.* 1998), are 'poor' predictors of behavior.[3] Oswald *et al.* (2013: 18) conclude that 'the IAT provides little insight into who will discriminate against whom, and provides no more insight than explicit measures of bias'. Many have taken Oswald and colleagues' conclusion to be definitive (especially many critics outside psychology; e.g., Bartlett 2017; Singal 2017; Yao and Reis-Dennis n.d.).

2.2. TEMPORAL INSTABILITY

Individuals' scores on implicit measures fluctuate considerably over time. Multiple longitudinal studies have demonstrated low correlations between individuals' scores on implicit measures across days, weeks, and months (Cooley and Payne 2017; Cunningham *et al.* 2001; Devine *et al.* 2012; Gawronski *et al.* 2017). This instability – a reflection of 'test–retest' reliability – is particularly pronounced on implicit measures of *racial* attitudes. Put simply, *ceteris paribus*, an individual's score on an implicit measure at T_1 – particularly an implicit measure of racial attitudes – is a weak predictor of her score on that same measure at T_2. Moreover, scores on implicit measures appear to be more temporally unstable than individuals' scores on corresponding explicit measures (Gawronski *et al.* 2017).

3. *Attitude–behavior relations*

In this section, we first provide a broader context for questions about using attitudes – whether measured implicitly or explicitly, directly or indirectly – to predict behavior. Then we apply these background points to the case of implicit bias.

3.1. BACKGROUND

Predicting the future is difficult. Scientists have achieved remarkable success in predicting outcomes in some domains, particularly with respect to highly controlled systems in the natural sciences (e.g., predicting the trajectories of celestial objects). In other domains, even in the natural sciences, predictive success is less impressive (e.g., meteorology). A crucial job for philosophers of science is to calibrate our expectations of the predictive abilities of scientific models to various kinds of phenomena. For example, *ceteris paribus*, we should expect less predictive accuracy of more complex and less controlled systems than simpler ones.

The following discussion is aimed at calibrating our expectations of implicit measures to the realities involved with predicting the behavior of individual human beings. What should count as predictive success? For context, compare the average correlations between individuals' scores on implicit measures and measures of behavior ($r = 0.14$ to $r = 0.28$) to correlation coefficients between other constructs and behavior: beliefs and stereotypes about outgroups and behavior ($r = 0.12$; Talaska *et al.* 2008); IQ and income (r = 0.2–0.3; Strenze 2007); SAT scores and freshman grades in college ($r = 0.24$; Wolfe and Johnson 1995); and parents' and their children's socioeconomic status (r = 0.2–0.3; Strenze 2007).[4] In general, it is rare for any well-known individual difference variable to approach the so-called large or medium-to-large zero-order correlations (i.e., $r \geq 0.4$) with meaningful behaviors and outcomes. Admittedly, a profound skeptic of psychological research, or a dyed-in-the-wool situationist, might simply conclude that *all* these individualistic measures are 'poor'.[5] In what follows, however, we will argue that there are independently plausible, theory-driven and data-driven grounds for *expecting* precisely these small positive average correlations between isolated psychological measures and behavior, when other important factors are ignored. While there are some exceptions to this pattern (discussed below, in this section), expectations of predictive success should be modest.[6] Indeed, we would worry about a massive 'file drawer' problem for research on implicit bias if the reported correlations between implicit measures and behavior exceeded these comparative norms.[7]

In a similar vein, it is crucial to recall that research on implicit measures partly arose out of the recognition that self-report (i.e., explicit) measures of attitudes predict behavior within this small to small-to-medium range as well, a predictive pattern that has been repeatedly confirmed in the more recent meta-analyses. This range is, therefore, no less and no more 'damning' for self-report than for implicit measures. Attitude researchers have not, nor should have, abandoned self-report measures, given these findings. Our point here is *not* that self-report measures are perfect and that implicit measures are just as good as them. Rather, our point is that while some proponents of implicit measures may have exaggerated their status as golden pipelines into the deep truth of individual's minds, these meta-analyses of measure–behavior correlations confirm that implicit measures fall within the range of other familiar and useful psychological tests. Self-report and implicit measures have distinctive strengths and weaknesses, which we discuss in the next section.

Since the 1970s, attitude researchers have recognized that the key question is not *whether* self-reported attitudes predict behavior just as such, but rather, *when* they predict behavior. One important lesson is that attitudes better predict behavior when there is clear correspondence between the attitude object and the behavior in question (Ajzen and Fishbein 1977). For example, while generic attitudes toward the environment do not predict recycling behavior very well, specific attitudes toward recycling do (Oskamp *et al.* 1991).[8] In the 1970s and 1980s, a consensus emerged that attitude–behavior relations depend in general on the particular behavior being measured, the conditions under which the behavior is performed, and the person who is performing the behavior (e.g., Zanna and Fazio 1982). A wealth of theoretical models of attitude–behavior relations takes these facts into account to make principled predictions about when attitudes do and do not predict behavior (e.g., Fazio 1990).

Indeed, stepping back from issues in psychometrics, the thought that any specific attitude will predict a range of behavior, regardless of behavior-specific, context-specific, and person-specific variables, conflicts with basic long-understood truisms about the mind. A person who likes hot dogs may be thought of as being disposed to eat hot dogs, but only when controlling for obvious variables. Does she believe that eating hot dogs is morally or religiously inappropriate? Is she dieting? Full from a big meal? Did she just floss? Is it 7:30 AM and simply an odd time to eat a hot dog? Are there other food options that she prefers nearby? Is she pretending to prefer escargot over hot dogs in order to impress a new acquaintance? Liking hot dogs, just as such, does not predict eating hot dogs in every, or even in the preponderance, of situations; we should expect low 'zero-order' correlations here. But concluding from this that self-reported liking of hot dogs is entirely useless for the prediction of hot dog-related behavior would be absurd. Behavior prediction depends on assessing people's attitudes in conjunction with their other attitudes and beliefs, their contexts, as well as with facts about the specific behavior in question.

Even attitudes that strongly correspond with behavior are only reliably predictive under theoretically expected conditions. Attitudes toward politicians, and toward political parties, tend to be relatively strongly associated with voting intentions and voting behavior, for example (for review, see Reyna *et al.* 2005). But Fazio and Williams (1986) found that the length of time it took participants to respond on a rating scale to questions about then-presidential candidates Ronald Reagan and Walter Mondale moderated the relationship between their attitudes and their actual voting behavior. Fazio and Williams characterized these response latencies as indicators of 'attitude accessibility'. For voters with highly accessible attitudes (i.e., those who responded quickly), 80% of the variance in their voting behavior was predicted by their attitudes toward Reagan and Mondale. For voters with low attitude accessibility (i.e., those who responded slowly), only 44% of the variance in their voting behavior was predicted by their attitudes toward Reagan and Mondale.[9]

When the behavior in question is socially sensitive, such as intergroup behavior involving racial attitudes, predicting it becomes even more difficult. Intergroup behavior – such as hiring decisions, interactions between police and civilians, and doctors' medical prescriptions – is inherently socially sensitive. Moreover, these kinds of intergroup behaviors are ambiguous in an important respect. In the sense that the attitude corresponding to eating hot dogs is liking hot dogs, what is the attitude corresponding to hiring more qualified men than qualified women for a job? Preferring men to women is a very rough proxy for this, as are related associations between men and, say, intelligence or competence. This ambiguity, along with the inherent difficulty of assessing people's attitudes in situations where they are frequently motivated to hide them, must frame any expectations of the attitude–behavior relationship.

3.2. IMPLICIT MEASURES AND BEHAVIORAL PREDICTION

The core upshot of the discussion thus far is that, without taking person-specific, context-specific, and behavior-specific moderators into account, a meta-analysis of any valid attitude measure ought to find consistent, positive, but low predictive relations between attitudes and behavior. *This is exactly what has been found in meta-analyses of implicit measures.* Not a single meta-analysis of implicit measures has reported nonsignificant correlations close to zero or negative correlations with behavior.

Some critics interpret the idea of moderated predictive relations as a sign of failure, suggesting that such arguments are post hoc attempts to save the field. It is true that many studies in this area have ignored the basic, theory-driven considerations that we are emphasizing (Machery, p.c.). But it is crucial to recognize that theories of key moderators and processes predate the current wave of criticism of implicit bias and have received little attention from critics. The idea of moderated prediction by implicit measures is at the core of many highly cited theories, including the MODE model ('motivation and opportunity as determinants'; Fazio and Towles-Schwen 1999), aversive racism theory (Dovidio and Gaertner 2004), the dual-attitude model (Wilson *et al.* 2000), and the reflective-impulsive model (Strack and Deutsch 2004). Moreover, the hypothesis of moderated prediction has been directly tested in the very first studies that used measures of implicit bias to predict behavior (i.e., Dovidio *et al.* 1997; Fazio *et al.* 1995). Expanding on extant theory and empirical findings, Friese *et al.* (2008) offered a systematic, detailed, theoretically guided review of when and why implicit measures do and do not predict behavior, identifying variables such as whether individuals were or were not motivated to control their spontaneous impulses, whether individuals were high or low in working memory capacity (and so were differentially *able* to control their impulses), and so on.

For example, consistent with predictions derived from the MODE model (Fazio 1990) and aversive racism theory (Gaertner and Dovidio 1986), some of the first studies on the prediction of behavior with implicit measures found that implicit bias showed stronger relations with spontaneous compared with deliberate behavior, whereas explicit bias showed stronger relations with deliberate compared with spontaneous behavior (Dovidio *et al.* 1997; Fazio *et al.* 1995). If the authors of these studies had focused on average correlations, they would have found positive, but relatively weak relations between implicit bias and behavior. However, such average correlations would conceal the insight that predictive relations should be high only for certain types of behavior (i.e., spontaneous versus deliberate behavior).[10]

Some meta-analyses have done an exemplary job of coding for key moderators. For example, Cameron *et al.* (2012) analyzed 167 studies that used sequential priming measures to predict behavior. They found a small average correlation between sequential priming scores and behavior ($r = 0.28$). Yet correlations were substantially higher under theoretically expected conditions ($r = 0.4$) and close to zero under conditions where no relation would be expected ($r = -0.004$). Cameron and colleagues identified their moderators from the fundaments of three influential dual-process models of social cognition.[11] While these models differ in important ways, they converge in predicting that implicit measures will correspond more strongly with behavior when agents have low motivation or low opportunity to engage in deliberation or when implicit associations and deliberatively considered propositions are consistent with each other. It is important to emphasize that Cameron *et al.* did not simply take the stated expectations of the authors of the included studies for granted in coding moderators. Rather, the dual-process moderators were derived *a priori* from the theoretical literature.

A more recent meta-analysis of intergroup IAT studies focuses on both theoretical and design-related factors that moderate relations between implicit measures and behavior. Kurdi *et al.* (2018) find an average zero-order correspondence of $r = 0.37$ in studies using the most effective IAT designs. Specifically, they find an average correspondence of $r = 0.37$ when they restrict their analysis to studies using a standard IAT rather than an IAT variant, like the 'Single-Category' IAT (Karpinski and Steinman 2006), a relative and graded measure of behavior (e.g., deciding precisely how much money to donate to a black student organization relative to a predominantly white student organization, rather than simply deciding whether to donate some fixed sum to a black student organization or not), and that have high correspondence between the attitude and behavioral measures (in the same vein that we discussed earlier, viz., recycling attitudes and recycling behavior). We discuss these findings, as well as potential ways to improve implicit measures, in Section 6. The point here is that, as follows from Cameron and colleagues' review, it is premature to conclude anything from average correlations between measures of attitudes and behaviors, which ignores any theoretical expectations of the relations between them.

The same points are key when the incremental validity of implicit measures is taken into consideration. That is, meta-analyses find that the IAT predicts behaviors over and above self-report measures (e.g., Kurdi *et al.* 2018). This does *not* mean that the IAT is superior (or inferior) to self-report measures. Rather, it means that the IAT *adds* to the predictive power of self-report measures. Moreover, some specific studies that find no predictive power in self-report measures find significant predictive power in corresponding implicit measures (e.g., Agerström and Rooth 2011). Kurdi *et al.* (2018) replicated this result using a modeling approach recommended by Westfall and Yarkoni (2016), showing that the incremental predictive validity of implicit and explicit measures is highly similar. This statistical approach controls more effectively for self-reported attitudes as well as for measurement error. What remains the key open question is when – in what domains, under what conditions, and so on – implicit measures outperform explicit measures and vice versa. Each makes distinctive contributions to the prediction of behavior. Moreover, the conditions under which one type of measure outperforms the other will most likely vary on theoretically expected grounds (e.g., when the topic is socially sensitive and when the motivation or opportunity to control spontaneous impulses is low).

It is likely that the best predictions will be achieved by combining both types of measure. For example, using a large dataset (N = 24,015), Bar-Anan and Vianello (2018) incorporated seven different implicit measures and three different explicit measures, on three distinct topics (race, politics, and the self), and found that a dual-construct model fits the data better than a single-construct model.[12] Indeed, even in the case of political attitudes, for which self-report measures are strongly predictive of political behavior, implicit measures have incremental validity. Friese *et al.* (2007) found that both self-reported attitudes toward political parties in Germany and self-reported intentions to vote strongly predicted voting behavior. Yet in both cases, a variant of the IAT – the single-target IAT – showed incremental predictive validity. Greenwald *et al.* (2009b) report similar findings in the US context using both self-reported and implicit race attitude measures to predict voting decisions for John McCain and Barack Obama. *Both* self-report and implicit measures predicted voting. This is noteworthy given the electoral power of 'undecided' voters who fail to report clear political preferences (Galdi *et al.* 2008).

4. *Temporal stability*

When relevant theoretical and methodological variables (e.g., theory-based moderators) are ignored and averaged over, one should expect the relevant attitude–behavior correlations to be positive, but relatively small overall. This expectation is borne out in the above meta-analyses. This finding does not necessarily impugn the validity or utility of the constructs posited by theories of attitudes or implicit social cognition.

Low test–retest stability in implicit measures represents a more serious challenge to their psychometric quality.[13] But here, too, attention to various *a priori* and theoretically derived considerations is crucial. In particular, the stability over time of a person's scores on implicit measures must be understood in terms of the interaction of individual differences with situational variables. Philosophers and social scientists have long debated the relative importance, for explaining human thought and action, of features of individuals (e.g., beliefs, traits, and virtues) versus situational variables (e.g., wealth, culture, and modes of production). Implicit bias presents a case study for the requirement of focusing on their complex, yet theoretically meaningful, ways of interacting.

4.1. BACKGROUND

Some measures of individual difference variables are more stable over time than others. For example, measures of intelligence and personality tend to be much more stable than implicit measures. From 2 to 12 weeks, for example, the Wechsler Adult Intelligence Scale (4th Edition) has test–retest reliability of 0.7–0.9 (Wechsler 2014). Estimates of test–retest reliability on implicit measures vary, as we will explain below, but they tend to be much lower (roughly from 0.3 to 0.55). How should we understand measures that are malleable across time and situations? Changes in scores across time on any measure that attempts to capture differences between individuals can be due to a number of different factors. If a scale is being used to track changes over time in a person's weight, a lower reading on second measurement could reflect that the person lost weight, is at a much higher elevation above sea level, or that the scale is broken. The first possibility explains the change in terms of the person; the second explains the change in terms of context; the third in terms of a faulty instrument. The dominant interpretation within the field of intergroup psychology has been that the instability of implicit measures across time indicates changes within persons, namely, malleability within their implicit associations. Some researchers, most notably Keith Payne and colleagues, have taken the second route, arguing that more attention ought to be paid to changes in situational factors. Critics of implicit bias research have taken the third route, arguing that test–retest instability suggests that measures like the IAT are faulty instruments.

Although we agree with some prominent criticisms related to the low stability of implicit measures (e.g., low temporal stability of a measure can undermine its capability to predict behavior over time), it is important to distinguish between temporal stability of a measure and its validity in capturing a particular construct, given that the measured construct may be unstable over time. Based on these considerations, we are sympathetic to a combination of the first and second interpretations. However, these interpretations are not easy to disambiguate, because changes across time on implicit measures may reflect relatively short-term changes in the momentary accessibility of stored information, given some change in the agent's situation, or longer-term changes in the structure or strength of a person's associations themselves (Gawronski and Bodenhausen 2006; Madva 2016a; see Section 5 for discussion of Payne and colleagues' 'Bias of Crowds' model). In the next section, we focus on reasons to think that, when relevant contextual features are held constant, implicit measures can capture more stable trait-like features of individuals. This conclusion gives reason to doubt the third interpretation, that instability in implicit measures across time shows these instruments to be faulty. But our point in the next section is not only that improvements in the design of implicit measures can lead to greater test–retest reliability. Our point is also that measuring the transient thoughts and feelings that people have in specific contexts is itself valuable both for explanatory and normative reasons. If tired people reliably show more bias on implicit measures than well-slept people, for example, then we will not only understand a feature of the dynamics of short-term changes in implicit bias, but also a potential element of mitigating bias (e.g., by instituting limits on the number of hours police officers can work in one stretch).

4.2. IMPLICIT ATTITUDES AND TEMPORAL STABILITY

In a recent longitudinal study, Gawronski and colleagues found that implicit measures of self-concept, political attitudes, and racial attitudes were less temporally stable across 1–2 months than corresponding explicit measures. It would, however, be premature to interpret such findings as evidence that implicit measures are unreliable, or generally less reliable or useful than explicit measures. For one, both the IAT and AMP are internally consistent by the standards used to evaluate explicit measures of attitudes (Gawronski *et al.* 2017). Internal consistency reflects the correlations between items on a scale. Measures that are internally consistent are thought to be measuring something systematic within individuals; *ceteris paribus*, low test–retest stability combined with adequate internal consistency suggests that the variability between individuals' scores at different times reflects the malleability and context sensitivity of personal characteristics, rather than flaws in the tools to measure them (Payne *et al.* 2017; see also Brownstein 2016; Brownstein and Madva 2012; Gawronski and Cesario 2013).[14] The natural analogies here are to measures of heart rate and blood pressure, which fluctuate dramatically across contexts (because the measures are accurately tracking that heart rate and blood pressure themselves fluctuate dramatically), but are also used to measure more chronic, trait-like features of individuals. Of course, using these tools to measure chronic constructs requires, among other things, doing as much as possible to hold fixed the contexts of measurement, hence the phrase '*resting* heart rate'. Strictly speaking, a one-time measurement of heart rate is capturing a fleeting event, but, with careful attention to context, it can be used to gather (partial, defeasible) evidence about more stable heart-rate dispositions.

Similarly, research suggests that people show temporally stable individual differences on implicit measures when there are meaningful contextual constraints and these constraints are held constant over time for all participants. In the absence of such constraints, scores on implicit measures are significantly shaped by incidental contextual factors that may differ from person to person, as well as over time, thereby producing low test–retest correlations. Gschwendner *et al.* (2008) offer an example that illustrates this insight. They assessed German participants' implicit evaluations of German versus Turkish faces on an IAT and varied the background context during each block of the text (i.e., they manipulated the blank space on the computer screen immediately below the target images and attribute words). Participants in the experimental condition saw a picture of a mosque, which is a conceptually meaningful context for evaluations of Muslims, while participants in the control condition saw a picture of a garden, which is conceptually irrelevant for evaluations of Muslims. Gschwendner and colleagues then compared stability of participant scores over a 2-week period. Whereas participants in the control condition showed a relatively low stability coefficient of 0.29, participants in the experimental condition showed a relatively high stability coefficient of 0.72. This latter correlation is notably similar to Gawronski and colleagues' overall finding for stability of explicit measures (r = 0.75). This is only one study, of course, and thus needs to be replicated and expanded upon. But it is suggestive that implicit measures are not unavoidably unstable. Rather, the conditions under which they are, and are not, stable must be better understood.

It bears emphasizing that research along these lines predates psychology's replication crisis and the competing meta-analyses of implicit measures. These studies were not driven by *post hoc* attempts to 'rescue' a dying research paradigm, but by a combination of empirical evidence and *a priori* and theory-based considerations about the relevance of contextual cues to patterns of concept accessibility and activation. Note, moreover, that Gschwendner and colleagues have effectively taken general hypotheses about the relevance of context and *built these insights into the measures* themselves, making the context *part* of the measure (think again of resting heart rate). This manipulation makes implicit measures less of a volatile, Rorschach-like indicator of the transient thoughts and activation patterns that happen to spontaneously cross an individual's mind at a given time, and more of an indicator of a stable, trait-like disposition (i.e., the disposition to respond with certain thoughts, feelings, and behavioral impulses in a certain range of contexts). (See also discussion in section 5.2 of Mischel and Shoda's (1995) comparable insights regarding personality research.)

There is additional suggestive evidence for relevant moderators of test–retest correlations elsewhere. Cooley and Payne (2017), for example, show significantly increased temporal stability in AMP scores when images of target groups, rather than images of target individuals, are used. Moreover, there appear to be important differences in the temporal stability of implicit associations with different contents. Gawronski and colleagues' finding of $r = 0.54$ is an average correlation across all the implicit measures they considered. For implicit political attitudes, however, they found a stability coefficient of 0.64 (when using an AMP to consider participants' relative implicit preferences for Trump or Clinton). The stability of participants' implicit racial attitudes on an AMP was decidedly lower – $r = 0.38$. An analogous situation is found in explicit attitude measures; the temporal stability of explicit political attitudes is significantly higher than the temporal stability of explicit racial attitudes. We note that conclusions drawn from these comparisons must be tentative, given the differences between measures that are not being held constant (e.g., stimulus materials). But we take these results to be suggestive.

We have described three factors that may affect the test–retest stability of implicit measures: the salience of relevant context cues; the type of images used as targets; and the content of the attitudes being measured. The broader lesson here is that there may be meaningful and temporally stable differences between individuals when there are meaningful contextual constraints. In the absence of such constraints, what is on a person's mind is influenced by incidental contexts and in ways that vary between individuals and over time. Theoretical frameworks, such as the associate-propositional evaluation model (Gawronski and Bodenhausen 2006, 2011), the situated inference model (Loersch and Payne 2011), and the resource computation model (Cesario and Jonas 2014), aim to predict these patterns. Our goal is not to defend any particular theoretical model, but rather to point to the data that any model must explain. The mere fact of low test–retest stability in implicit measures, considered in independence of any of these data and the theories that predict them, is not sufficient to cast implicit bias research wholesale into doubt. Our hope is that continued research in the vein of Gschwendner *et al.* and Cooley and Payne, which has already indicated promising avenues for increasing the stability of implicit measures over time, may contribute to the ability of these measures to capture more trait-like than state-like characteristics of individuals.

5. *Additional worries*

5.1. HYPE

The hype surrounding research on implicit bias is a cause for concern, because the lofty presentation of implicit measures in public discussion sometimes goes far beyond (and sometimes conflict with) the empirical evidence. Researchers themselves are, in some cases, guilty of this, although the most egregious cases are found in the popular press,[15] diversity consulting firms,[16] and the websites of academic departments.[17] Such mischaracterizations can lead to serious problems: the misuse of money intended to combat discrimination, the creation of misguided public policy, and popular misunderstanding of the workings of science. Overhype can also contribute to cultural skepticism about scientific knowledge and expertise, for example, when initially well-publicized results do not survive replication, and it can feed back into research itself, in the sense that scholars may be incentivized to advertise flashy findings before they are sufficiently well supported. Researchers seeking to understand the social epistemology of science ought to consider implicit bias as a case study of the problems associated with hype at the nexus of social science, philosophy, and science communication.

Ultimately, however, there are two issues here, not one. Our focus is on the scientific standing of research on implicit bias. The problems associated with overhyping this research may have problematic social effects, and may also reciprocally cause problems for what researchers choose to investigate and so on, but this no more means the two issues are one than the Pope's excommunication of Galileo, which surely negatively affected astronomy, meant anything for the truth or falsity of heliocentrism. Critics should not impugn the research itself by pointing to the ways in which journalists, corporations, politicians, and some researchers have misunderstood it, any more than they should impugn climate science on the grounds that only 48% of Americans believe that global climate change is caused by human activity,[18] or impugn the theory ('just a theory!') of evolution on the grounds that only 19% of Americans believe that 'human beings developed over millions of years, but God had no part in this process'.[19] Such examples should make salient the many serious challenges facing contemporary science communication and education, and perhaps implicit bias researchers could benefit from more explicit training on this front.[20]

There are some easy fixes, in our view, to improve the popular understanding of implicit bias. For example, Greenwald *et al.* (2015) caution against using the IAT as a diagnostic tool for classifying kinds of people (e.g., as 'implicit racists'). We agree. This does not mean that the IAT is not a legitimate measure of meaningful differences between individuals or between participants assigned to different experimental conditions (see discussion in Section 5.2); individual differences are not typologies, and *racism* is a hugely complex and loaded label. For this reason, it may be that the format of the feedback given on Project Implicit ought to be revised. At present, test-takers of the race IAT are told that their scores indicate 'slight', 'moderate', or 'strong' bias. One could take this to mean that they are slight, moderate, or strong 'implicit racists'. But no single instance of measurement, using one tool, can determine how biased a person is *tout court*.

Finally, it must be recognized that there remain genuine, ongoing disagreements among those who take these measures seriously, about the nature of the underlying psychological constructs, the best ways to measure them, and so on. Are the attitudes measured by the IAT consciously accessible? Are they propositionally structured? Researchers disagree, but they may gloss over these debates while trying to communicate novel findings that are not directly related to these questions.[21] In any case, these communicative shortcomings are a stand-alone issue – which is hardly unique to implicit bias – and ought not be confused with the scientific legitimacy of the research itself.

5.2. SITUATIONISM

Payne and colleagues call for a shift away from the individual differences approach to understanding implicit bias, toward an approach that prioritizes situational contexts. This is a welcome advance. In short, their 'bias of crowds' model treats these instruments more as measures of situations than of persons. This model is meant to explain five common findings: (1) Average group-level scores of implicit bias are very robust and stable; (2) children's average scores of implicit measures are nearly identical to adults' average scores, suggesting little aggregate change over time; (3) aggregate levels of implicit bias at the population level (e.g., regions, states, and countries) are both highly stable and strongly associated with discriminatory outcomes and group-based disparities; yet, as we discussed in Sections 2–4, (4) individual differences in implicit bias have small-to-medium zero-order correlations with discriminatory behavior; and (5) individual test–retest reliability is low over weeks and months. Regarding (3), for example, Payne and colleagues used Project Implicit data to analyze average levels of implicit racial bias for each of the US states, finding that, from 1 year to the next, the test–retest stability is quite high ($r = 0.76$) and remains so even over a 10-year span ($r = 0.69$). Moreover, a slew of recent studies have found that these regional average scores correlate with real-world outcomes. Even after adjusting for variables such as explicit bias, residential segregation, and local levels of violent crime and unemployment, Hehman *et al.* (2017) find that greater racial disparities in police shootings in metropolitan regions of the USA are associated with higher levels of implicit racial bias in those regions ($\beta = 0.39$). Findings like this – and there are numerous others (e.g., Leitner *et al.* 2016, 2018; Marini *et al.* 2013; Orchard and Price 2017; Rae *et al.*

2015) – underscore the need for careful study of the relations between implicit bias and social situations and structures.[22]

But how could implicit measures be so powerful at the group level, as in (1)–(3), while so volatile at the individual level, as in (4) and (5)? The bias of crowds model accounts for the stark differences between individual-level and group-level data by appealing to the 'accessibility' of social concepts in individuals' minds, that is, the 'likelihood that a thought, evaluation, stereotype, trait, or other piece of information' becomes activated and poised to influence behavior. Payne and colleagues argue that concept accessibility varies primarily and dramatically as a function of the situation the individual is in. By analogy, one might predict, for example, that the color green will not generally make thoughts of beer highly accessible, except around St. Patrick's Day. Most research on implicit intergroup bias over the past two decades has focused on the differences *between* individuals in concept accessibility (e.g., by contrasting the behavior of individuals who do versus do not automatically associate 'Black' with 'weapon'), but Payne and colleagues propose that researchers focus anew on the situational causes of concept activation (e.g., contrasting the situations that do versus do not activate Black–weapon associations). 'Although concept accessibility can, in principle, vary both chronically and situationally, there is little empirical evidence for chronic accessibility that gives rise to stable individual differences in implicit intergroup bias' (236), they write. 'Instead, most of the systematic variance in implicit biases appears to operate at the level of situations' (236).

As we emphasized earlier, we embrace the call for a renewed emphasis on situational moderators of the accessibility of the concepts underlying implicit bias. Recognizing this does not signal the death of the individual differences approach, however.[23] In seeing why, a comparison with research in personality psychology is instructive. Despite the binary uptake in recent philosophical discussion, which pits 'persons' versus 'situations' (e.g., Harman 1999), it is a defining assumption of foundational theories that personality only emerges *in interaction with* situational variables (e.g., Bandura 1978; Lewin 1936; Mischel and Shoda 1995; see Cervone *et al.* (2001) for discussion). In the most general sense, the interactionist view states that personality consists of differences between how individuals react to situations, rather than general, context-free individual differences (Fleeson 2004; see also Doris' (2002) account of 'local traits'). Evidence for this view is that personality variables (e.g., 'extroversion') are weak predictors of how people will behave in any one given situation but are strongly correlated with behavioral trends over time (Fleeson 2004). This is strikingly similar to the evidence Payne and colleagues marshal in favor of their bias of crowds model; implicit measures are weak predictors of how people will behave in any one given situation, but are strongly associated with aggregated data.[24]

What the person *versus* situation debate obscures, in both personality research and implicit bias research, is that predictions ought to be derived primarily from theoretical models of person-by-situation interactions. In their reply to critics, Payne and colleagues posit concept accessibility as the mechanism linking systemic (i.e., situation-based) biases to cognitive processes. Theoretical predictions of concept accessibility via person-by-situation activation are many. Samayoa and Fazio (2017) point to attitude strength, for example. Stronger attitudes are associated with more powerful person-based effects; weaker attitudes are associated with more powerful situation-based effects. Gawronski and Bodenhausen (2006, 2011) point toward many more, most notably the way in which the same stimulus can activate different concepts for individuals, given the structure and strength of their mental associations. While Payne and colleagues disagree with these researchers over the comparative emphasis that should be placed on situational versus personal effects, all involved accept a view of implicit bias in terms of person-by-situation interactions, and none assert that research on individual differences is dead.

Much of these differences in approach can also be understood in terms of differing explananda (Kurdi and Banaji 2017). Population-level research, like Hehman and colleagues', treats individual differences and short-term temporal instability in measurements as error. Here, the object of study *is* the aggregate itself.[25] In contrast, traditional implicit bias research treats individual differences and short-term changes as the objects of study and treats peripheral situational variables as noise. Person-by-situation interactions are a third object of study. For example, Cesario *et al.* (2010) studied personality-by-implicit-bias-by-situation interactions. They found that, among participants who automatically associated 'black' with 'danger', 'black' activated flight-related concepts in the context of an open field and fight-related concepts in the context of an enclosed booth. Moreover, they found that while implicitly biased participants with non-confrontational personalities tended to sit farther away from a black interlocutor (i.e., avoiding potential confrontations with a member of a perceived-dangerous group), implicitly biased participants with confrontational personalities tended to sit closer.

Of course, when statistical analyses become complex in these ways, there are familiar risks associated with generating hypotheses after the results are known ('HARKing') and with mining the data until a particular effect reaches statistical significance ('p-hacking'). But the same precautionary steps and best practices that are widely recommended to avoid these missteps are straightforwardly applicable in implicit bias research as well (e.g., preregister studies, specify the number of participants in advance based on power estimates, and introduce more stringent tests of statistical significance) and ought to be applied to implicit bias research. Implicit bias research is neither more nor less vulnerable to problems like p-hacking than are other fields of empirical study (within psychology and beyond), and we offer no novel solutions to address these problems here (but see, e.g., Loersch and Payne 2011 and Cesario and Jonas 2014 for discussions of how contextual moderators like those we discuss here should inform replication research).

Finally, we note the connection between these issues and calls within philosophy and the social sciences for greater attention to the structural causes of inequalities and discrimination. Several theorists have been critical of implicit bias research for its putatively individualistic focus (Banks and Ford 2008, Dixon *et al.* 2012, Haslanger 2015), and we are sympathetic with the general point that the field has focused on the contents of participants' minds to the exclusion of contexts, norms, and social structures. However, the findings assembled by Payne and colleagues suggest that there is nothing inherently individualistic in the measures themselves. To the contrary, aggregate IAT scores evidently represent an alternative strategy for assessing systemic and structural discrimination. Several other studies have used variants of the IAT to detect implicit perceptions of social norms and regularities (Yoshida *et al.* 2012; Peach, Yoshida, and Zanna 2011; Peach, Yoshida, Spencer, Zanna, and Steele 2011; Walton *et al.* 2015; cf. Brownstein and Madva 2012), and we support these directions for future research.

5.3. CORRELATIONS BETWEEN IMPLICIT MEASURES

Some critics have argued that low correlations between different implicit measures are a cause for concern (e.g., Machery 2016, 2017a,b). If a set of measures are valid representations of the same construct, then they should, *ceteris paribus*, correlate with one another. Our view is that (a) some measures are simply more reliable than others, at least for certain purposes, and (b) none of these measures is 'process-pure', which is to say that different measures 'tap into' different processes in theoretically expected ways.[26] Neither of these points undermines research on implicit bias.

Several reviews have found that not all implicit measures are equally reliable (Bar-Anan and Nosek 2014; Gawronski and De Houwer 2014; Payne and Lundberg 2014). In general, the IAT and the AMP tend to do best in terms of their internal consistencies. If this is so, then one should not expect a reliable measure to correlate with measures with weaker psychometric properties. Now, it is the case that even well-validated measures – variations of the IAT and the AMP – do not always strongly correlate. One explanation for this is that correlations may vary as a function of the content of what is being measured (e.g., self-esteem, race, or political evaluations), with the weakest correlations found when the most complex concepts are targeted. Moreover, it is difficult to control for differences in content. Consider an AMP targeting attitudes toward homosexuality and an IAT targeting associations between homosexuality and competence. It is not clear that the affective feelings elicited by pictures of (for example) gay men kissing represents the same concepts as those elicited by the presentation of pairings of words associated with gay men and words associated with competence.

This example leads to our second point. Even if the target attitudes are controlled across measures, it is well established that each of these measures is influenced by a range of automatic and controlled processes, such that different measures capture different components of individual performance, including motivation and self-regulatory capacity, in addition to 'pure' concept accessibility.[27] For example, the IAT measures implicit bias in terms of participants' relative speed or accuracy in categorizing pairings of concepts, whereas the AMP measures neither speed nor accuracy and instead treats bias in terms of participants' intentional judgments (misattributions) about the pleasantness of stimuli (for a discussion, see Gawronski and De Houwer 2014). Given the AMP's slower pace and reliance on untimed deliberate judgments, we find unsurprising the recent evidence suggesting that the AMP is more closely related to explicit measures than it is to other implicit measures (Bar-Anan and Nosek 2014; Bar-Anan and Vianello 2018; cf. Payne and Lundberg 2014). Ultimately, the relatively low correlations between implicit measures are not so much an anomaly that threatens the field as it is a pedestrian empirical finding, which has begun to be explored (e.g., Bar-Anan and Nosek 2016; Moran *et al.* 2017). We would certainly welcome further theory-based and experimental investigation into the mechanisms explaining performance on these distinct measures (cf. Bishara and Payne 2009; Conrey *et al.* 2005; Payne *et al.* 2010), which should illuminate when and why they come apart and, in turn, which specific measures are most apt for which specific aims, contexts, psychological processes, and behavioral predictions.

6. *Future directions*

There is significant room to improve implicit measures. Here, we briefly note some areas of promise.

In response to criticism, IAT researchers in particular have often pointed to an 'accumulation' model of discrimination and social disparities (e.g., Greenwald *et al.* 2015).[28] For example, Greenwald *et al.* (2015) identify two conditions under which a tool that measures statistically small effects can track behavioral patterns with large social significance. One is when the effects apply to many people and the other is when the effects are repeatedly applied to the same person. Following Messick (1995), Greenwald and colleagues refer to this as the 'consequential validity' of a measure. They provide the following example to show how small effects that apply to many people can be significant for predicting discrimination:

> As a hypothetical example, assume that a race IAT measure has been administered to the officers in a large city police department, and that this IAT measure is found to correlate with a measure of issuing citations more frequently to Black than to White drivers or pedestrians

(profiling). To estimate the magnitude of variation in profiling explained by that correlation, it is necessary to have an estimate of variability in police profiling behavior. The estimate of variability used in this analysis came from a published study of profiling in New York City (Office of the Attorney General, 1999), which reported that, across 76 precincts, police stopped an average of 38.2% (SD = 38.4%) more of each precinct's Black population than of its White population. Using [Oswald *et al.*' (2013)] $r = .148$ value as the IAT–profiling correlation generates the expectation that, if all police officers were at 1 SD below the IAT mean, the city-wide Black–White difference in stops would be reduced by 9,976 per year (5.7% of total number of stops) relative to the situation if all police officers were at 1 SD above the mean. Use of [Greenwald *et al.*' (2009a)] larger estimate of $r = .236$ increases this estimate to 15,908 (9.1% of city-wide total stops).This suggests that a measure with a correlational effect size of 0.236 (or even 0.148) has a role to play in understanding patterns of discriminatory behavior. So too is this the lesson when discriminatory impact accumulates over time by repeatedly affecting the same person (e.g., in hiring, testing, healthcare experiences, and law enforcement). With repetition, even tiny impact increases the chances of significantly undesirable outcomes. Greenwald *et al.* (2015) draw an analogy to a large clinical trial of the effect of aspirin in preventing heart attacks:

> The trial was terminated early because data analysis had revealed an unexpected effect for which the correlational effect size was the sub-small value of $r = .035$. This was 'a significant ($P < 0.00001$) reduction [from 2.16% to 1.27%] in the risk of total myocardial infarction [heart attack] among those in the aspirin group' (Steering Committee of the Physicians' Health Study Research Group, 1989). Applying the study's estimated risk reduction of 44% to the 2010 U.S. Census estimate of about 46 million male U.S. residents 50 or older, regular small doses of aspirin should prevent approximately 420,000 heart attacks during a 5-year period.

The effect of taking aspirin on the likelihood of having a heart attack for any particular person is tiny, but the sub-small value of the effect was significant enough to terminate data analysis in order to advance the research for use in public policy.

Our defense of implicit measures has *not* relied on arguments about accumulation mechanisms like these. While we think these models are promising – particularly in light of recent studies correlating population-level IAT data with real-world inequities (e.g., Hehman *et al.* 2017, discussed earlier) – we recognize that they are only, at present, statistical models. While this does not mean that they are worthless, future research must vindicate this approach using data from implicit measures themselves.[29]

In addition to pursuing the model of accumulation mechanisms (Mallon 2017; Lombrozo and Mallon 2017), we believe there is significant room for methodological improvement in the measurement of individuals' implicit attitudes. For example, as we noted earlier, Kurdi *et al.* (2018) find that methodological design drastically affects IAT correlations with criterion measures. They recommend the use of the standard IAT (rather than its variants) with high polarity between attributes; relative measures of behavior; and strong correspondence between attitudes and criterion behavior. In this vein, Cooley and Payne (2017) find that the AMP showed greater within-individual test–retest reliability when it used photos of groups of people rather than isolated individuals. This tweak, which might also benefit the IAT, improves the likelihood that the measure is truly tapping into attitudes about *groups* rather than about idiosyncratic features of particular individuals or photos that are not directly related to the construct being measured.

Madva and Brownstein (2018) have also made specific proposals for improving the IAT by, for example, targeting the activation of specific associations in specific contexts with specific behavioral outcomes. For example, Levinson, Smith, and Young (2014) developed a novel IAT that found a tendency to associate white faces with words like ‘merit’ and ‘value’ and black faces with words like ‘expendable’ and ‘worthless’. This measure predicted, among other things, that mock jurors with stronger ‘white-value/black-worthless’ associations were more likely to sentence a black defendant to death rather than life in prison. *Prima facie*, this correlation suggests that the race-value IAT is tracking, at least to some extent, something like a disposition to devalue black lives. This suggestion is supported by the fact that another IAT that measured associations between white and black faces and words like ‘lazy’ and ‘unemployed’ did not predict death sentencing. These measures capture different implicit associations and should predict different behavior. Of course, these are stand-alone studies that need to be replicated. The point is not that these studies necessarily reveal the truth. Rather, the point is that these measures are successful – if their apparent success is ultimately vindicated – by targeting specific, contextually relevant associations in theoretically informed ways. In this vein, we call for more theoretical and empirical work on how specific contexts activate specific associations and behavior.

All that said, more tweaks of this kind will only take implicit measures so far. The mind is populated with many different types of attitudes, biases, concepts, and cognitive structures, each of which will be better poised to explain distinctive spheres of social judgment and action. The assumption that *all* social biases will be best measured either by feeling thermometers or by timed concept-accessibility tasks like the IAT is empirically and theoretically unwarranted. Consider, for example, research on generics (Leslie *et al.* 2015), on motivated propositional reasoning due to cognitive dissonance and consistency (Gawronski and Strack 2012) and moral and political values (Jost 2017; Tetlock *et al.* 2000), on ‘fast and frugal’ heuristics and biases (Gigerenzer 2008; Hewstone, Benn, and Wilson 1988; Kahneman 2011; Peer and Gamliel 2013), on the dependency relations in networks of concepts (Meyer *et al.* 2015; del Pinal, Madva, and Reuter 2017; Sloman, Love, and Ahn 1998), and on the tradition of research on ‘schemas’ that preceded the turn to concept accessibility and semantic priming in implicit social cognition (see, e.g., Valian 1998 for a review). All of these psychological constructs may be relevant to explaining and predicting contemporary prejudice and discrimination, but many of them will elude detection on the sorts of self-report questionnaires and timed implicit measures that have come to dominate the field. All of these constructs will also interact with each other, as well as with contextual variables. For example, one well-established moderator in the heuristics and biases literature is mood (Chartrand *et al.* 2006). In short, people in good moods rely more on 'fast' heuristic processes while those in bad moods engage in slower and more deliberate cognitive elaboration. The same pattern evidently applies to implicit biases. Participants in good moods are more likely to make judgments based on their implicit biases, while those in bad moods are more likely to make judgments in line with their reported attitudes (Forgas 2011; Holland *et al.* 2012). More research on moderators like these is needed (cf. Madva 2018).

7. *Conclusion*

Critics may interpret our arguments as an attempt to draw a rosy picture, suggesting that all is well with research on implicit bias. That was *not* the goal of this article. Our goal was to show that extant concerns have very different implications when the criticism is considered in the broader context of research on attitudes and implicit measures. To be sure, such a perspective raises important questions about a common narrative in the field, according to which implicit biases reflect stable traits that cause discriminatory behavior in an unconditional manner. This narrative is difficult to defend in light of the empirical evidence. However, this conclusion does not imply that implicit measures are useless and should be abandoned. As we explained in this article, implicit measures are better understood as reflecting what is on a person's mind in a given moment, which is shaped by complex interactions of person-related and situation-related factors. Incorporating attention to these factors in future research promises to improve behavioral prediction, test–retest reliability, and our broader understanding of larger-scale social phenomena related to health, discrimination, and inequality. Interpreted in this manner, implicit measures are still invaluable tools for understanding the workings of the human mind.

Michael Brownstein
Department of Philosophy
John Jay College and The Graduate Center, CUNY

Alex Madva
Department of Philosophy
Cal Poly Pomona

Bertram Gawronski
Department of Psychology
University of Texas at Austin

NOTES

[1] http://philosophyofbrains.com/2017/01/17/how-can-we-measure-implicit-bias-a-brains-blog-roundtable.aspx.

[2] But see Section 5 for additional sources of criticism. Implicit measures have been criticized for various reasons that we do not discuss, as well (e.g., coding of IAT scores; low internal consistency of evaluative priming tasks), but most of these concerns are related to specific instruments, and therefore do not question the field of implicit bias research in general, because the task-specific shortcomings of certain measures are compensated by the strengths of others and vice versa (Gawronski and De Houwer 2014).

[3] The IAT is a reaction-time measure that asks participants to sort words and pictures into categories as fast as possible while making as few mistakes as possible. An IAT score is computed by comparing participants' speed and accuracy on trials in which they must sort the words and pictures in ways that are consistent with common social stereotypes and prejudices with trials in which they must sort the words and pictures in ways that are inconsistent with common social stereotypes and prejudices. Participants' speed and accuracy on these sorting tasks are thought to indicate the strength of their associations (between, for example, black faces and negative words); in turn, the strength of their associations is thought to contribute to thought, feeling, and action. Other implicit measures, such as the Affect Misattribution Procedure (AMP) (Payne et al. 2005; discussed below and in Section 2.2), prime participants with stimuli (e.g., an image of a black face, presented for a fraction of a second, in the case of the AMP) and then assess the effects of the prime on similar measures of thought, feeling, and action. For more in-depth explanation of the IAT, AMP, and other implicit measures, see Brownstein (2015) and Gawronski and De Houwer (2014).

[4] See also, for example, Poropat (2009) and Richardson *et al.*'s (2012) meta-analyses of the correlations between GPA and an array of psychological and other constructs, such as the Big Five personality traits, intelligence, goals, and demographic variables like age, sex, and socioeconomic status. According to the College Board's own experiments on predictors of undergraduate GPA, the 'raw' uncorrected correlations are $r = 0.27$ for high school GPA, 0.3 for the most recent revision of the math section of the SAT, and 0.33 for the reading and writing section of the SAT (Shaw *et al.* 2016). Combining the math and verbal sections brings the correlation up to 0.35, and then adding high school GPA brings it up to 0.4. In a comparative survey of individual difference measures predicting meaningful life outcomes including mortality and divorce, Roberts *et al.* (2007: 337) write, 'the effect sizes associated with SES, cognitive abilities, and personality traits were all uniformly small-to-medium in size. This finding is entirely consistent with those from other reviews showing that most psychological constructs have effect sizes in the range between .10 and .40 on a correlational scale.' For a classic survey of correlations illustrating the rarity of surpassing the 0.4 threshold, within psychology and beyond, see Meyer *et al.* (2001), and for a more recent survey of research on the roles of cognitive ability and expertise in predicting job performance, see Hambrick *et al.* (2019).

[5] We specifically address situationism in Section 5.

[6] Machery (p.c.) points to IQ as a measure that reaches large zero-order correlations with behavior. Although some meta-analyses have claimed to show this, the studies and meta-analytic techniques to support them are at least controversial. See, for example, Richardson and Norgate's (2015) analysis of the often-reported claim that IQ correlates 0.5 with job performance. We also note that the independent predictive power of any one genuinely non-redundant psychological construct (e.g., personality traits, IQ, chronic goals, explicit beliefs, and implicit

biases) is necessarily constrained by the predictive powers of all the other non-redundant constructs – and, given the mind's complexity, there are quite a few such constructs to go around (cf. Section 6).

[7] Kurdi *et al.* (2018) find little evidence of publication bias, using several tests.

[8] More recently, Axt (2018) assembled a large body of evidence from Project Implicit that suggests that many explicit measures of racial attitudes suffer by virtue of being too indirect, for example, by measuring attitudes toward affirmative action as a proxy for attitudes toward African Americans. While these indirect self-report measures may be less likely to be influenced by participants' self-presentation concerns, they may introduce noise by virtue of measuring beliefs and attitudes not directly related to race.

[9] We are unaware of any attempts to replicate this study. We are not endorsing its findings but using it to illustrate our point.

[10] But see Cameron *et al.* (2012) and Kurdi *et al.* (2018) for further analysis of the idea that implicit measures show stronger relations with spontaneous behavior and explicit measures show stronger relations with deliberate behavior. Our point is not to defend this point about spontaneous versus deliberate behavior *per se* but to illustrate how the informational value of meta-analyses depends upon its treatment of key moderators.

[11] Specifically, from MODE model (Fazio 1990), 'associative-propositional evaluation' model (Gawronski and Bodenhausen 2006), and 'meta-cognitive model' (Petty, Briñol, and DeMarree 2007).

[12] Although the implicit measures were *partly* related to explicit measures (cf. Nosek 2007), there was significant shared variance across the implicit measures that (with the partial exception of the AMP) was *not* shared with the explicit measures, which strongly suggests that the distinct implicit measures are, to varying degrees and in perhaps varying ways, tapping into some single 'implicit' construct (whether that construct is a process, representation, evaluation, etc.). See also Cunningham *et al.* (2001) and Schimmack (Forthcoming).

[13] We do not mean to suggest that test–retest reliability is a more important psychometric consideration than predictive validity in general. Rather, as we have argued, the low predictive validity of implicit measures is to be expected when considered in independence of theoretical expectations; in contrast, the temporal instability of implicit measures is a much deeper challenge, because it questions their validity as measures of trait-like constructs and their suitability for predicting behavior over time.

[14] Variance on any given measures can be divided into (1) systematic construct variance; (2) systematic measurement error; and (3) random error. Both systematic construct variance and systematic measurement error contribute to internal consistency. Bar-Anan and Vianello (2018) use a multitrait, multimethod approach to, among other things, begin to disentangle the roles of (1), (2), and (3) in implicit and explicit measures. Note, however, that for any given manipulation that causes a change on an implicit measure, the effect could be related more to (2) than (1), that is, changing the score without changing the construct of interest (by analogy, think of concerns in education about merely 'teaching to the test').

[15] In the *New York Times*, Nicholas Kristof writes, 'It's sobering to discover that whatever you believe intellectually, you are biased about race, gender, age or disability.' See <https://www.nytimes.com/2015/05/07/opinion/nicholas-kristof-our-biased-brains.html>. As we discussed earlier, explicit beliefs about social concepts are, in fact, strong moderators of implicit attitudes about those concepts.

[16] In their document 'Proven Strategies for Addressing Unconscious Bias in the Workplace', a company called CDO Insights offers the following: 'Each one of us has some groups with which we consciously feel uncomfortable, even as we castigate others for feeling uncomfortable with our own groups. These conscious patterns of discrimination are problematic, but, again, they pale in comparison to the unconscious patterns that impact us every day.' <http://www.cookross.com/docs/UnconsciousBias.pdf>. There is little reason to think, though, that the

problems associated with explicit bias 'pale in comparison' with the problems associated with implicit bias. See Section 5.4.

[17] The Diversity and Cultural Competence website of the Johns Hopkins Medical School asserts that, for example, 'The IAT has demonstrated to be both reliable and valid at detecting an individual's level of implicit bias.' See <www.hopkinsmedicine.org/odcc/implicit_association_test.html>. This seems to suggest that the IAT is a valid diagnostic tool of individuals' 'level' of bias. But this is not true and reflects a misunderstanding of the nature of statistical averages.

[18] http://www.pewinternet.org/2016/10/04/public-views-on-climate-change-and-climate-scientists/

[19] http://news.gallup.com/poll/210956/belief-creationist-view-humans-new-low.aspx

[20] Cf. http://www.aldakavlilearningcenter.org/

[21] These tendencies are even more understandable in the political context of communicating about implicit bias. When journalists or politicians use research on implicit bias to suggest very broadly that we are all implicated in structures of injustice, politically and culturally motivated critics sometimes make a concerted effort to portray these statements as criticisms of the individual character of ordinary Americans, police officers, and so on.

[22] See Hehman *et al*. (2019) for discussion of pertinent psychometric questions about these studies and their novel approach to the use of implicit measures.

[23] Payne and colleagues themselves do not advocate an end to the individual differences approach, although some commenters attribute this view to them.

[24] In the case of implicit bias, the data are aggregated between persons, as in Hehman and colleagues' research. In the case of personality measures, the data are aggregated within persons over time, at least in Fleeson's influential research. See Machery 2017a for discussion. But see also Rentfrow *et al*. (2015), for example, for research on between-individual, regional differences in personality traits, and see Madva (2016c) for empirical and normative discussion of individual-level factors and concept accessibility.

[25] A question for future research is how to determine the appropriate aggregate. On the basis of what theoretical expectations should researchers identify the town, city, state, or nation as the aggregate?

[26] For example, although measures like the IAT are typically advertised as a pipeline into automatic processes, an extensive theoretical and experimental literature demonstrates that performance on these measures is influenced, to some extent, by more deliberate, controlled processes, and a variety of data-analytic models and tools have emerged to shed light on the comparative contributions of these different types of psychological processes (e.g., Conrey *et al*. 2005). There is also evidence that much of the *apparent* variance between implicit measures is simply another manifestation of the contextual variability of concept accessibility, that is, the same phenomenon that explains the low test–retest reliability *within* implicit measures. For example, Cunningham *et al*. (2001) found much higher correlations between the IAT and the evaluative priming task after controlling for measurement error.

[27] See, for instance, analyses of various implicit measures using multinomial models (e.g., Bishara and Payne 2009; Conrey *et al*. 2005; Payne *et al*. 2010).

[28] See also Valian (1998, 2005) and Sripada's comment at <http://philosophyofbrains.com/2017/01/17/how-can-we-measure-implicit-bias-a-brains-blog-roundtable.aspx>.

[29] See critical discussion of this example in Oswald *et al*. (2015), who argue that inferences about police officers cannot be drawn given that the distribution of IAT scores for police officers is unknown. This strikes us as unpersuasive, given that Greenwald and colleagues present the example explicitly as hypothetical and there is little reason to think that police officers would demonstrate *less* anti-black bias on the IAT compared with the average IAT population pool. (See Mekawi and Bresin (2015) for a meta-analysis of related shooter-bias studies.) Moreover, Greenwald and colleagues' general point about small effect sizes having significant consequences has been made elsewhere, irrespective of the details of this particular example. Rosenthal (1991), for example, (Rosenthal and Rubin 1982) shows that an r of 0.32 for a cancer treatment,

compared with placebo, which accounts for only 10% of variance, translates into a survival rate of 66% in the treatment group compared with 34% in the placebo group. That being said, we note another caveat about the 'accumulation mechanisms' defense of implicit bias research. The statistical model of Greenwald *et al*. (2015) is based on the assumption of additivity, and there are reasons to assume a multiplicative instead of an additive model. In a multiplicative model, 'trickle down effects' become less impactful within a causal chain of factors, because the probabilities of implicit bias influencing outcomes would have to be multiplied for each step of the causal chain. For example, in a causal chain including two mediating variables that may be influenced by implicit bias and one societal outcome as a distal variable, the likelihoods for each step would have to be multiplied to assess the impact of implicit bias on the outcome. Thus, even if implicit bias explains 20% of variance for each step of the causal chain, it ultimately explains only 4% of variance in the societal outcome.

REFERENCES

Agerström, J. and Rooth, D. O. (2011). 'The Role of Automatic Obesity Stereotypes in Real Hiring Discrimination,' *Journal of Applied Psychology* 96, pp. 790–805.

Ajzen, I. and Fishbein, M. (1977). 'Attitude–Behavior Relations: A Theoretical Analysis and Review of Empirical Research,' *Psychol Bull* 84, pp. 888–918.

Axt, J. R. (2018). 'The Best Way to Measure Explicit Racial Attitudes Is to Ask about Them,' *Social Psychological and Personality Science* 9(8), pp. 896–906.

Bandura, A. (1978). 'The Self System in Reciprocal Determinism,' *American Psychologist* 33, pp. 344–358.

Banks, R. R. and Ford, R. T. (2008). '(How) Does Unconscious Bias Matter: Law, Politics, and Racial Inequality,' *Emory LJ* 58(5), pp. 1053–1122.

Bar-Anan, Y. and Nosek, B. (2014). 'A Comparative Investigation of Seven Implicit Attitude Measures,' *Behavioral Research* 46, pp. 668–688.

Bar-Anan, Y., & Nosek, B. A. (2016). 'A Comparison of the Sensitivity of Four Indirect Evaluation Measures to Evaluative Information.' https://osf.io/n3j6r/

Bar-Anan, Y. and Vianello, M. (2018). 'A Multi-method Multi-trait Test of the Dual-attitude Perspective,' *J Exp Psychol Gen* 147(8), pp. 1264–1272. https://doi.org/10.1037/xge0000383.

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

Bishara, A. J. and Payne, B. K. (2009). 'Multinomial Process Tree Models of Control and Automaticity in Weapon Misidentification,' *J Exp Soc Psychol* 45(3), pp. 524–534.

Brownstein, M. (2015/2019). 'Implicit Bias.' *The Stanford Encyclopedia of Philosophy*. Zalta, E. (Ed.) <http://plato.stanford.edu/entries/implicit-bias/>.

Brownstein, M. (2016). 'Context and the Ethics of Implicit Bias.' In M. Brownstein & J. Saul (Eds.), *Implicit Bias and Philosophy, Volume 2 (pp. 215–234)*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198766179.003.0010

Brownstein, M. (Forthcoming). 'Skepticism about Bias,' in E. Beeghly and A. Madva (eds) *An Introduction to Implicit Bias: Knowledge, Justice, and the Social Mind*. Routledge.

Brownstein, M. and Madva, A. (2012). 'The Normativity of Automaticity,' *Mind and Language* 27(4), pp. 410–434.

Buckwalter, W. (2018). 'Implicit Attitudes and the Ability Argument,' *Philosophical Studies*. https://doi.org/10.1007/s11098-018-1159-7.

Cameron, C. D., Brown-Iannuzzi, J. L. and Payne, B. K. (2012). 'Sequential Priming Measures of Implicit Social Cognition: A Meta-analysis of Associations with Behavior and Explicit Attitudes,' *Pers Soc Psychol Rev* 16, pp. 330–350.

Cervone, D., Shadel, W. and Jencius, S. (2001). 'Social-cognitive Theory of Personality Assessment,' *Pers Soc Psychol Rev* 5(1), pp. 33–51.

Cesario, J., Plaks, J., Hagiwara, N., Navarret, C. D. and Higgins, E. T. (2010). 'The Ecology of Automaticity: How Situational Contingencies Shape Action Semantics and Social Behavior,' *Psychol Sci* 21(9), pp. 1311–1317.

Cesario, J. and Jonas, K. J. (2014). 'Replicability and Models of Priming: What a Resource Computation Framework Can Tell Us about Expectations of Replicability,' *Social Cognition* 32(Supplement), pp. 124–136. https://doi.org/10.1521/soco.2014.32.supp.124.

Chartrand, T. L., van Baaren, R. B. and Bargh, J. A. (2006). 'Linking Automatic Evaluation to Mood and Information Processing Style: Consequences for Experienced Affect, Impression Formation and Stereotyping,' *J Exp PsycholGen* 135(1), pp. 70–77.

Conrey, F., Sherman, J., Gawronski, B., Hugenberg, K. and Groom, C. (2005). 'Separating Multiple Processes in Implicit Social Cognition: The Quad-model of Implicit Task Performance,' *J Pers Soc Psychol* 89, pp. 469–487.

Cooley, E. and Payne, B. K. (2017). 'Using Groups to Measure Intergroup Prejudice,' *Pers Soc Psychol Bull* 43(1), pp. 46–59.

Cunningham, W. A., Preacher, K. J. and Banaji, M. R. (2001). 'Implicit Attitude Measures: Consistency, Stability, and Convergent Validity,' *Psychol Sci* 12(2), pp. 163–170. https://doi.org/10.1111/1467-9280.00328.

Del Pinal, G., Madva, A. and Reuter, K. (2017). 'Stereotypes, Conceptual Centrality and Gender Bias: An Empirical Investigation,' *Ratio* 30(4), pp. 384–410.

Devine, P., Forscher, P., Austin, A. and Cox, W. (2012). 'Long-term Reduction in Implicit Race Bias: A Prejudice Habit-breaking Intervention,' *J Exp Soc Psychol* 48(6), pp. 1267–1278.

Dixon, J., Levine, M., Reicher, S. and Durrheim, K. (2012). 'Beyond Prejudice: Are Negative Evaluations the Problem and Is Getting Us to Like One Another More the Solution?' *Behavioral and Brain Sciences* 35(6), pp. 411–425.

Doris, J. (2002). *Lack of Character: Personality and Moral Behavior*. Cambridge: Cambridge University Press.

Dovidio, J. F. and Gaertner, S. L. (2004). 'Aversive Racism,' *Advances in Experimental Social Psychology* 36, pp. 1–52.

Dovidio, J. F., Kawakami, K., Johnson, C., Johnson, B. and Howard, A. (1997). 'On the Nature of Prejudice: Automatic and Controlled Components,' *J Exp Soc Psychol* 33, pp. 510–540.

Fazio, R. and Williams, C. (1986). 'Attitude Accessibility as a Moderator of the Attitude–Perception and Attitude–Behavior Relations: An Investigation of the 1984 Presidential Election,' *J Pers Soc Psychol* 51(3), pp. 505–514.

Fazio, R. H. (1990). 'Multiple Processes by which Attitudes Guide Behavior: The MODE Model as an Integrative Framework,' *Advances in Experimental Social Psychology* 23, pp. 75–109.

Fazio, R. H., Jackson, J. R., Dunton, B. C. and Williams, C. J. (1995). 'Variability in Automatic Activation as an Unobtrusive Measure of Racial Attitudes: A Bona Fide Pipeline?' *J Pers Soc Psychol* 69, pp. 1013–1027.

Fazio, R. H. and Towles-Schwen, T. (1999). 'The MODE Model of Attitude–Behavior Processes,' in S. Chaiken and Y. Trope (eds) *Dual Process Theories in Social Psychology*. New York: Guilford, pp. 97–116.

Fleeson, W. (2004). 'Moving Personality beyond the Person–Situation Debate,' *Curr Dir Psychol Sci* 13(2), pp. 83–87.

Forgas, J. P. (2011). 'She Just Doesn't Look like a Philosopher …? Affective Influences on the Halo Effect in Impression Formation,' *Eur J Soc Psychol* 41(7), pp. 812–817.

Forscher, P., Lai, C., Axt, J., Ebersole, C., Herman, M., Devine, P. and Nosek, B. (2019). 'A Meta-analysis of Change in Implicit Bias,' *J Pers Soc Psychol* 117(3), pp. 522–559. https://doi.org/10.1037/pspa0000160.

Friese, M., Bluemke, M. and Wänke, M. (2007). 'Predicting Voting Behavior with Implicit Attitude Measures: The 2002 German Parliamentary Election,' 54(4), pp. 247–255. https://doi.org/10.1027/1618-3169.54.4.247.

Friese, M., Hofmann, W. and Schmitt, M. (2008). 'When and Why Do Implicit Measures Predict Behaviour? Empirical Evidence for the Moderating Role of Opportunity, Motivation, and Process Reliance,' *Eur Rev Soc Psychol* 19(1), pp. 285–338. https://doi.org/10.1080/10463280802556958.

Gaertner, S. L. and Dovidio, J. F. (1986). 'The Aversive Form of Racism,' in J. F. Dovidio and S. L. Gaertner (eds) *Prejudice, Discrimination, and Racism*. Orlando, FL: Academic Press, pp. 61–89.

Galdi, S., Arcuri, L. and Gawronski, B. (2008). 'Automatic Mental Associations Predict Future Choices of Undecided Decision-makers,' *Science* 321, pp. 1100–1102.

Gawronski, B. and Bodenhausen, G. V. (2006). 'Associative and Propositional Processes in Evaluation: An Integrative Review of Implicit and Explicit Attitude Change,' *Psychol Bull* 132, pp. 692–731.

Gawronski, B. and Bodenhausen, G. V. (2011). 'The Associative-propositional Evaluation Model. Theory, Evidence, and Open Questions,' *Advances in Experimental Social Psychology* 44, pp. 59–127.

Gawronski, B. and Bodenhausen, G. V. (2017). 'Beyond Persons and Situations: An Interactionist Approach to Understanding Implicit Bias,' *Psychological Inquiry* 28(4), pp. 268–272.

Gawronski, B. and Cesario, J. (2013). 'Of Mice and Men: What Animal Research Can Tell Us about Context Effects on Automatic Response in Humans,' *Pers Soc Psychol Rev* 17(2), pp. 187–215.

Gawronski, B. and De Houwer, J. (2014). 'Implicit Measures in Social and Personality Psychology,' in H. T. Reis and C. M. Judd (eds) *Handbook of Research Methods in Social and Personality Psychology*, 2nd edn. New York, NY: Cambridge University Press, pp. 283–310.

Gawronski, B., Morrison, M., Phills, C. and Galdi, S. (2017). 'Temporal Stability of Implicit and Explicit Measures: A Longitudinal Analysis,' *Pers Soc Psychol Bull* 43(3), pp. 300–312.

Gawronski, B. and Strack, F. (eds) (2012). *Cognitive Consistency: A Fundamental Principle in Social Cognition*, 1st edn. New York, NY: The Guilford Press.

Gigerenzer, G. (2008). *Gut Feelings: The Intelligence of the Unconscious*, Reprint edn. London: Penguin Books.

Greenwald, A., McGhee, D. and Schwartz, J. (1998). 'Measuring Individual Differences in Implicit Cognition: The Implicit Association Test,' *J Pers Soc Psychol* 74, pp. 1464–1480.

Greenwald, A., Poehlman, T., Uhlmann, E. and Banaji, M. (2009a). 'Understanding and Using the Implicit Association Test: III Meta-analysis of Predictive Validity,' *J Pers Soc Psychol* 97 (1), pp. 17–41.

Greenwald, A., Smith, C., Sriram, N., Bar-Anan, Y., and Nosek, B. (2009b). 'Implicit Race Attitudes Predicted Vote in the 2008 U.S. Presidential Election', *Analysis of Social Issues and Public Policy* 9(1), pp. 241–253

Greenwald, A. G., Banaji, M. R. and Nosek, B. A. (2015). 'Statistically Small Effects of the Implicit Association Test Can Have Societally Large Effects,' *J Pers Soc Psychol* 108, pp. 553–561.

Gschwendner, T., Hofmann, W. and Schmitt, M. (2008). 'Differential Stability: The Effects of Acute and Chronic Construct Accessibility on the Temporal Stability of the Implicit Association Test,' *Journal of Individual Differences* 29, pp. 70–79.

Hambrick, D. Z., Burgoyne, A. P., & Oswald, F. L. (2019). 'Domain-general Models of Expertise: The Role of Cognitive Ability.' In D. Z. Hambrick, A. P. Burgoyne, & F. L. Oswald, P. Ward, J. Maarten Schraagen, J. Gore, & E. Roth (Eds.), *The Oxford Handbook of Expertise*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780198795872.013.3

Harman, G. (1999). 'Moral Philosophy Meets Social Psychology: Virtue Ethics and the Fundamental Attribution Error,' *Proceedings of the Aristotelian Society* 99, pp. 315–331.

Haslanger, S. (2015). 'Social Structure, Narrative, and Explanation,' *Canadian Journal of Philosophy* 45(1), pp. 1–15.

Hehman, E., Calanchini, J., Flake, J. K. and Leitner, J. B. (2019). 'Establishing Construct Validity Evidence for Regional Measures of Explicit and Implicit Racial Bias,' *J Exp Psychol Gen* 148(6), pp. 1022–1040. https://doi.org/10.1037/xge0000623.

Hehman, E., Flake, J. K. and Calanchini, J. (2017). 'Disproportionate Use of Lethal Force in Policing Is Associated with Regional Racial Biases of Residents,' *Social Psychological and Personality Science*, 1948550617711229.

Hermanson, S. (2017a). 'Implicit Bias, Stereotype Threat, and Political Correctness in Philosophy,' *Philosophies* 2(2), p. 12. https://doi.org/10.3390/philosophies2020012.

Hermanson, S. (2017b). in M. Brownstein and J. Saul (eds) *Review of Implicit Bias and Philosophy*, Volume 1 & 2. Oxford University Press, 2016, pp. 315–322.

Hermanson, S. (2018). 'Rethinking Implicit Bias: I Want My Money Back.' http://leiterreports.typepad.com/blog/2018/04/sean-hermanson-rethinking-implicit-bias-i-want-my-money-back.html

Hewstone, M., Benn, W. and Wilson, A. (1988). 'Bias in the Use of Base Rates: Racial Prejudice in Decision-making,' *Eur J Soc Psychol* 18(2), pp. 161–176. https://doi.org/10.1002/ejsp.2420180207.

Holland, R. W., de Vries, M., Hermsen, B. and van Knippenberg, A. (2012). 'Mood and the Attitude–Behavior Link: The Happy Act on Impulse, the Sad Think Twice,' *Social Psychological and Personality Science* 3(3), pp. 356–364.

Jost, J. T. (2017, June). 'A Theory of System Justification.' *Psychological Science Agenda*. Retrieved from http://www.apa.org/science/about/psa/2017/06/system-justification.aspx

Kahneman, D. (2011). *Thinking, Fast and Slow*. New York: Farrar, Straus and Giroux.

Karpinski, A. and Steinman, R. B. (2006). 'The Single Category Implicit Association Test as a Measure of Implicit Social Cognition,' *J Pers Soc Psychol* 91(1), p. 16.

Kurdi, B. and Banaji, M. R. (2017). 'Reports of the Death of the Individual Difference Approach to Implicit Social Cognition May Be Greatly Exaggerated: A Commentary on Payne, Vuletich, and Lundberg,' *Psychological Inquiry* 28(4), pp. 281–287.

Kurdi, B., Seitchik, A., Axt, J., Carroll, T., Karapetyan, A., Kaushik, N., Tomezsko, D., Greenwald, A. and Banaji, M. (2018). 'Relationship between the Implicit Association Test and Intergroup Behavior: A Meta-analysis,' *American Psychologist*. https://doi.org/10.1037/amp0000364.

Leitner, J. B., Hehman, E., Ayduk, O. and Mendoza-Denton, R. (2016). 'Racial Bias Is Associated with Ingroup Death Rate for Blacks and Whites: Insights from Project Implicit,' *Soc Sci Med* 170, pp. 220–227.

Leitner, J. B., Hehman, E. and Snowden, L. (2018). 'States Higher in Racial Bias Spend Less on Disabled Medicaid Enrollees,' *Soc Sci Med*. https://doi.org/10.1016/j.socscimed.2018.01.013.

Leslie, S.-J., Cimpian, A., Meyer, M. and Freeland, E. (2015). 'Expectations of Brilliance Underlie Gender Distributions across Academic Disciplines,' *Science* 347(6219), pp. 262–265. https://doi.org/10.1126/science.1261375.

Levinson, J. D., Smith, R. J. and Young, D. M. (2014). 'Devaluing Death: An Empirical Study of Implicit Racial Bias on Jury-eligible Citizens in Six Death Penalty States,' *NYUL Rev* 89, pp. 513–581.

Lewin, K. (1936). 'A Dynamic Theory of Personality: Selected Papers,' *J Nerv Ment Dis* 84(5), pp. 612–613.

Loersch, C. and Payne, B. K. (2011). 'The Situated Inference Model: An Integrative Account of the Effects of Primes on Perception, Behavior, and Motivation,' *Perspectives on Psychological Science* 6(3), pp. 234–252. https://doi.org/10.1177/1745691611406921.

Lombrozo, T., & Mallon, R. (2017, July 24). 'How Small Inequities Lead to Big Inequalities.' 13.7: Cosmos & Culture Blog. Retrieved from http://www.npr.org/sections/13.7/2017/07/24/539010535/how-small-inequities-lead-to-big-inequalities

Mac Donald, 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. Oxford, UK: Oxford University Press, pp. 104–129.

Machery, E. (2017a). 'Do Indirect Measures of Biases Measure Traits or Situations?' *Psychological Inquiry* 28(4), pp. 288–291.

Machery, E. (2017b). '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

Madva, A. (2016a). 'Why Implicit Attitudes Are (Probably) Not Beliefs,' *Synthese* 193(8), pp. 2659–2684. https://doi.org/10.1007/s11229-015-0874-2.

Madva, A. (2016b). 'A Plea for Anti-anti-individualism: How Oversimple Psychology Misleads Social Policy,' *Ergo, an Open Access Journal of Philosophy* 3(27), pp. 701–728. https://doi.org/10.3998/ergo.12405314.0003.027.

Madva, A. (2016c). 'Virtue, Social Knowledge, and Implicit Bias,' in M. Brownstein and J. Saul (eds) *Implicit Bias and Philosophy: Metaphysics and Epistemology: Volume 1*. Oxford: Oxford University Press, pp. 191–215.

Madva, A. (2017). 'Biased against Debiasing: On the Role of (Institutionally Sponsored) Self-transformation in the Struggle against Prejudice,' *Ergo, an Open Access Journal of Philosophy* 4(6), pp. 145–179.

Madva, A. (2018). 'Implicit Bias, Moods, and Moral Responsibility,' *Pacific Philosophical Quarterly*. https://doi.org/10.1111/papq.12212.

Madva, A. and Brownstein, M. (2018). 'Stereotypes, Prejudice, and the Taxonomy of the Implicit Social Mind,' *Noûs* 52(3), pp. 611–644.

Mallon, R. (2017). 'Psychology, Accumulation Mechanisms, and Race.' Presented at the the Society for Philosophy and Psychology 43nd Annual Meeting, Johns Hopkins University.

Marini, M., Sriram, N., Schnabel, K., Maliszewski, N., Devos, T., Ekehammar, B. et al. (2013). 'Overweight People Have Low Levels of Implicit Weight Bias, but Overweight Nations Have High Levels of Implicit Weight Bias,' *PLoS ONE* 8(12), e83543. https://doi.org/10.1371/journal.pone.0083543.

Mekawi, Y. and Bresin, K. (2015). 'Is the Evidence from Racial Bias Shooting Task Studies a Smoking Gun? Results from a Meta-analysis,' *J Exp Soc Psychol* 61, pp. 120–130.

Messick, S. (1995). 'Validity of Psychological Assessment: Validation of Inferences from Persons' Responses and Performances as Scientific Inquiry into Score Meaning,' *American Psychologist* 50, pp. 41–749.

Meyer, M., Cimpian, A. and Leslie, S.-J. (2015). 'Women are Underrepresented in Fields where Success Is Believed to Require Brilliance,' *Frontiers in Psychology* 6. https://doi.org/10.3389/fpsyg.2015.00235.

Mischel, W. and Shoda, Y. (1995). 'A Cognitive-affective System Theory of Personality: Reconceptualizing Situations, Dispositions, Dynamics, and Invariance in Personality Structure,' *Psychol Rev* 102(2), p. 246.

Moran, T., Bar-Anan, Y. and Nosek, B. A. (2017). 'The Effect of the Validity of Co-occurrence on Automatic and Deliberate Evaluations,' *Eur J Soc Psychol*. https://doi.org/10.1002/ejsp.2266.

Nosek, B. A. (2007). 'Implicit–Explicit Relations,' *Curr Dir Psychol Sci* 16(2), pp. 65–69. https://doi.org/10.1111/j.1467-8721.2007.00477.x.

Orchard, J. and Price, J. (2017). 'County-level Racial Prejudice and the Black–White Gap in Infact Health Outcomes,' *Soc Sci Med* 181, pp. 191–198.

Oskamp, S., Harrington, M. J., Edwards, T. C., Sherwood, D. L., Okuda, S. M. and Swanson, D. C. (1991). 'Factors Influencing Household Recycling Behavior,' *Environ Behav* 23(4), pp. 494–519.

Oswald, F. L., Mitchell, G., Blanton, H., Jaccard, J. and Tetlock, P. E. (2013). 'Predicting Ethnic and Racial Discrimination: A Meta-analysis of IAT Criterion Studies,' *J Pers Soc Psychol* 105, pp. 171–192.

Payne, B. K., Cheng, C. M., Govorun, O. and Stewart, B. D. (2005). 'An Inkblot for Attitudes: Affect Misattribution as Implicit Measurement,' *J Pers Soc Psychol* 89(3), pp. 277–293.

Payne, B. K., Hall, D. L., Cameron, C. D. and Bishara, A. J. (2010). 'A Process Model of Affect Misattribution,' *Pers Soc Psychol Bull* 36(10), pp. 1397–1408. https://doi.org/10.1177/0146167210383440.

Payne, B. K., Vuletich, H. A. and Lundberg, K. B. (2017). 'Flipping the Script on Implicit Bias Research with the Bias of Crowds,' *Psychological Inquiry* 28(4), pp. 306–311.

Payne, K. and Lundberg, K. (2014). 'The Affect Misattribution Procedure: Ten Years of Evidence on Reliability, Validity, and Mechanisms,' *Social and Personality Psychology Compass* 8(12), pp. 672–686. https://doi.org/10.1111/spc3.12148.

Peach, J. M., Yoshida, E., Spencer, S. J., Zanna, M. P. and Steele, J. R. (2011). 'Recognizing Discrimination Explicitly while Denying It Implicitly: Implicit Social Identity Protection,' *J Exp Soc Psychol* 47(2), pp. 283–292. https://doi.org/10.1016/j.jesp.2010.09.007.

Peach, J. M., Yoshida, E. and Zanna, M. P. (2011). 'Learning What Most People Like: How Implicit Attitudes and Normative Evaluations Are Shaped by Motivation and Culture and Influence Meaningful Behavior,' in J. P. Forgas, J. Cooper, W. D. Crano and S. J. Spencer (eds) *The Psychology of Attitudes and Attitude Change*. Philadelphia: Psychology Press, pp. 95–108.

Peer, E. and Gamliel, E. (2013). 'Heuristics and Biases in Judicial Decisions,' *Court Review* 49, pp. 114–119.

Petty, R. E., Briñol, P. and DeMarree, K. G. (2007). 'The Meta-cognitive Model (MCM) of Attitudes: Implications for Attitude Measurement, Change, and Strength,' *Social Cognition* 25, pp. 657–686.

Poropat, A. E. (2009). 'A Meta-analysis of the Five-factor Model of Personality and Academic Performance,' *Psychol Bull* 135(2), pp. 322–338. https://doi.org/10.1037/a0014996.

Rae, J. R., Newheiser, A. and Olson, K. R. (2015). 'Exposure to Racial Out-groups and Implicit Race Bias in the United States,' *Social Psychological and Personality Science* 6, pp. 535–543.

Rentfrow, P. J., Jokela, M. and Lamb, M. E. (2015). 'Regional Personality Differences in Great Britain,' *PLOS ONE* 10(3), e0122245. https://doi.org/10.1371/journal.pone.0122245.

Reyna, C., Henry, P. J., Korfmacher, W. and Tucker, A. (2005). 'Examining the Principles in Principled Conservatism: The Role of Responsibility Stereotypes as Cues for Deservingness in Racial Policy Decisions,' *J Pers Soc Psychol* 90(1), pp. 109–128.

Richardson, K. and Norgate, S. (2015). 'Does IQ Really Predict Job Performance?' *Applied Developmental Science* 19(3), pp. 153–169.

Richardson, M., Abraham, C. and Bond, R. (2012). 'Psychological Correlates of University Students' Academic Performance: A Systematic Review and Meta-analysis,' *Psychol Bull* 138, pp. 353–387. https://doi.org/10.1037/a0026838.

Roberts, B. W., Kuncel, N. R., Shiner, R., Caspi, A. and Goldberg, L. R. (2007). 'The Power of Personality: The Comparative Validity of Personality Traits, Socioeconomic Status, and Cognitive Ability for Predicting Important Life Outcomes,' *Perspectives on Psychological Science* 2(4), pp. 313–345. https://doi.org/10.1111/j.1745-6916.2007.00047.x.

Rosenthal, R. (1991). *Meta-analytic Procedures for Social Research*, Rev. edn. Newbury Park, CA: Sage.

Rosenthal, R. and Rubin, D. B. (1982). 'A Simple, General Purpose Display of Magnitude of Experimental Effect,' *J Educ Psychol* 74, pp. 166–169.

Samayoa, G. J. A. and Fazio, R. H. (2017). 'Who Starts the Wave? Let's Not Forget the Role of the Individual,' *Psychological Inquiry* 28(4), pp. 273–277.

Schimmack, Ulrich. Forthcoming. 'The Implicit Association Test: A Method in Search of a Construct.' https://osf.io/ajf9z/

Shaw, E. J., Marini, J. P., Beard, J., Shmueli, D., Young, L., & Ng, H. (2016). 'The Redesigned SAT® Pilot Predictive Validity Study: A First Look.' College Board. https://eric.ed.gov/?id=ED563472

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

Sloman, S. A., Love, B. C. and Ahn, W.-K. (1998). 'Feature Centrality and Conceptual Coherence,' *Cognit Sci* 22(2), pp. 189–228. https://doi.org/10.1207/s15516709cog2202_2.

Strack, F. and Deutsch, R. (2004). 'Reflective and Impulsive Determinants of Social Behavior,' *Pers Soc Psychol Rev* 8, pp. 220–247.

Strenze, T. (2007). 'Intelligence and Socioeconomic Success: A Meta-analytic Review of Longitudinal Research,' *Dermatol Int* 35, pp. 401–426.

Talaska, C., Fiske, S. and Chaiken, S. (2008). 'Legitimating Racial Discrimination: Emotions, Not Beliefs, Best Predict Discrimination in a Meta-analysis,' *Social Justice Research* 21(3), pp. 263–296.

Tetlock, P. E., Kristel, O. V., Elson, S. B., Green, M. C. and Lerner, J. S. (2000). 'The Psychology of the Unthinkable: Taboo Trade-offs, Forbidden Base Rates, and Heretical Counterfactuals,' *J Pers Soc Psychol* 78(5), pp. 853–870. https://doi.org/10.1037/0022-3514.78.5.853.

Valian, V. (1998). *Why So Slow? The Advancement of Women*. Cambridge, MA: MIT Press.

Valian, V. (2005). 'Beyond Gender Schemas: Improving the Advancement of Women in Academia,' *Hypatia* 20, pp. 198–213.

Walton, G. M., Logel, C., Peach, J. M., Spencer, S. J. and Zanna, M. P. (2015). 'Two Brief Interventions to Mitigate a 'Chilly Climate' Transform Women's Experience, Relationships, and Achievement in Engineering,' *J Educ Psychol* 107(2), pp. 468–485.

Wechsler, D. (2014). *Wechsler Adult Intelligence Scale–Fourth Edition (WAIS–IV)*. San Antonio, Texas: Psychological Corporation.

Westfall, J. and Yarkoni, T. (2016). 'Statistically Controlling for Confounding Constructs Is Harder than You Think,' *PLoS ONE* 11(3), pp. e0152719–e0152722. https://doi.org/10.1371/journal.pone.0152719.

Wilson, T. D., Lindsey, S. and Schooler, T. Y. (2000). 'A Model of Dual Attitudes,' *Psychol Rev* 107, pp. 101–126.

Wolfe, R. and Johnson, S. (1995). 'Personality as a Predictor of College Performance,' *Education and Psychological Measurement* 55(2), pp. 177–185.

Yao, V. & Reis-Dennis, S. n.d. Manuscript. 'I Love Women: The Conceptual Inadequacy of Implicit Bias.' http://peasoup.us/2017/09/love-women-conceptual-inadequacy-implicit-bias-yao-reis-dennis

Yoshida, E., Peach, J. M., Zanna, M. P. and Spencer, S. J. (2012). 'Not All Automatic Associations Are Created Equal: How Implicit Normative Evaluations Are Distinct from Implicit Attitudes and Uniquely Predict Meaningful Behavior,' *J Exp Soc Psychol* 48(3), pp. 694–706. https://doi.org/10.1016/j.jesp.2011.09.013.

Zanna, M. P. and Fazio, R. H. (1982). 'The Attitude–Behavior Relation: Moving toward a Third Generation of Research,' in M. P. Zanna, E. T. Higgins and C. P. Herman (eds) *Consistency in Social Behavior: The Ontario Symposium*, Volume 2. Hillsdale, N.J: Erlbaum, pp. 283–301.