DRAFT
11.13.17

Understanding Implicit Bias: How the Critics Miss the Point

Michael Brownstein, Alex Madva, and Bertram Gawronski

1. Introduction

What is the status of research on implicit bias? Criticism is ubiquitous. Recent meta-analytic reviews suggest that the Implicit Association Test 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, Lai, et al., ms). Prominent philosophers have questioned the validity of research on implicit social cognition altogether. Edouard Machery (2017), 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 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" (MacDonald 2017).

We argue that, while there is ample room for improvement, the current backlash against research on implicit bias is rooted in a superficial understanding of fundamental theoretical, empirical, and methodological issues in research on attitudes and implicit measures more broadly. To the extent that these issues are taken into account, research on the causes, structure, and behavioral effects of implicit bias clearly deserves 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 (§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 (§§3-4). We respond to potential criticism (§5), and then, finally, point to directions for future research (§6).

A quick note: our focus is on psychometrics. Philosophers, legal theorists, activists, and other social scientists have raised a number of important critical questions about research on implicit

---

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

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 (e.g., Brownstein 2016; Gawronski and Bodenhausen 2017; Madva 2016b) 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 $r = .14$ to $r = .28$ (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 vs. 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 attitudes account for between 1% and 8% of variance in intergroup discrimination. From these data, critics have concluded that implicit measures, in particular, the Implicit Association Test (IAT; Greenwald et al. 1998), are "poor" predictors of behavior. Oswald and colleagues conclude that "the IAT provides little insight into who will discriminate against whom, and provides no more insight than explicit measures of bias" (2013, 18). Many have taken Oswald and colleagues' conclusion to be definitive (especially many critics outside psychology; e.g., Bartlett 2017; Singal 2017; Yao & Reis-Dennis ms).

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 (Cunningham et al. 2001; Devine et al. 2012; Gawronski et al. 2017). Other implicit measures, such as the Affective Misattribution Procedure (AMP; Payne et al 2005), have yielded similar results (Cooley & Payne 2016) 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

[2] But see §5 for additional sources of criticism.

attitudes—is a weak predictor of her score on that same measure at T_2. Moreover, in certain contexts, such 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

Critics infer from these two sets of findings that implicit measures do not provide meaningful information about individuals' minds or about how individuals will behave in relevant intergroup situations. Although this conclusion may seem intuitively plausible, it rests on a number of premature background assumptions. In fact, as we explain in the remainder of this article, many of these background assumptions are wrong. In fact, we would argue that the entire framing of the debate is wrong. In this section, we first provide crucial background 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 behavior is difficult. Compare the average correlations between individuals' scores on implicit measures and measures of behavior (r = .14 to r = .28) to correlation coefficients between other constructs and behavior: beliefs and stereotypes about outgroups and behavior (r = .12; Talaska et al. 2008); IQ and income (r = .2-.3; Strenze 2007); SAT scores and freshman grades in college (r = .24; Wolfe and Johnson 1995); parents' socioeconomic status and your own socioeconomic status (r = .2-.3; Strenze 2007).[3] We would worry about a massive "file drawer" problem for research on implicit social cognition if the reported correlations between implicit measures and behavior exceeded these comparative norms.[4] In fact, research on implicit social cognition 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 (which are, therefore, no less "damning" for self-report than for implicit measures). This does not mean, however, that attitude researchers have, or should have, abandoned self-report measures. Rather, since the 1970s,

---

[3] See also, for example, Poporat (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. In general, it seems exceedingly rare for any well-known individual-difference variable to approach so-called "large" or "medium-to-large" zero-order correlations (i.e., $r \geq .4$) with meaningful behaviors and outcomes, such as GPA. Admittedly, a more profound skeptic of psychological research, or a dyed-in-the-wool situationist, might simply conclude that *all* these individualistic measures are "poor." In what follows, however, we will argue that there are independently plausible, theory- and data-driven grounds for *expecting* precisely these small positive average correlations between isolated psychological measures and behavior, when other important factors are ignored. We note, moreover, that the independent predictive power of any one genuinely non-redundant psychological construct (e.g., personality traits, IQ, chronic goals, explicit beliefs, implicit biases, etc.) 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. §6). We specifically address situationism in §5.

[4] Indeed, Kurdi et al. (ms) found little evidence of publication bias, using several tests.

researchers have recognized that the key question is not *whether* self-reported attitudes predict behavior, 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 & 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).[5] 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 & Fazio 1982). A wealth of theoretical models of attitude-behavior relations take 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 7:30AM 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 liking hot dogs is *irrelevant* to predicting 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. Liking hot dogs is more likely to correspond with eating hot dogs than with eating bratwurst, even though the two behaviors are similar.

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 Likert scale to questions about then-presidential candidate Ronald Reagan strongly 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. For voters with low attitude

---

[5] More recently, Axt (in press) assembled a large body of evidence from Project Implicit which 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 measures of explicit attitudes 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.

accessibility (i.e., those who responded slowly), only 44% of the variance in their voting behavior was predicted by their attitudes toward Reagan.

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 Attitudes and Behavioral Prediction

To claim that implicit measures should only predict behavior under specific conditions does not reflect a *post hoc* attempt to rescue implicit bias research from peril. Well before the replication crisis and the emergence of competing meta-analyses, Friese, Hofmann, and Schmitt (2008) offered systematic, detailed, theoretically derived, and empirically supported predictions about precisely when and why implicit measures should and should not predict behavior, 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. Meta-analyses that ignore such theoretically derived moderators should, therefore, find significant positive, yet small predictive relations between attitudes (whether self-reported or indirectly measured) and behavior. And this is exactly what has been found; not a single meta-analysis of implicit measures has reported nonsignificant correlations close to zero or negative correlations with behavior. Unfortunately, recent high-profile meta-analyses have largely ignored the relevant moderators of implicit attitude-behavior relations. Oswald and colleagues' (2013) meta-analysis included any study in which a race or ethnicity IAT and a behavioral outcome measure were used, but they did not differentiate between behavioral outcomes that should or should not be predicted on theoretical grounds. For example, Amodio and Devine (2006) found that the evaluative race-IAT predicted how much white participants desired to befriend a black student, but it did not predict how white participants expected a black student to perform on sports and academic trivia tasks. They also found that a stereotyping IAT, which measured associations between black and white faces and words associated with athleticism and intelligence, predicted how white participants would expect a black student to perform on sports and academic trivia tasks, but failed to predict white students' desire to befriend a black student. Amodio and Devine predicted these results on the basis of a theoretical model distinguishing between "implicit stereotypes" and "implicit prejudice."[6] If Amodio and Devine had reported the average correlation between their IATs and behavior, they would have found the same weak relationship reported by Oswald et al. (2013). In fact, the average correlation would have been even lower with r = .06. However, this average correlation would conceal the insight that predictive relations should be high only for theoretically "matching" types of behavior. Although this insight is widely accepted in the broader fields of attitudes and implicit social cognition, it has been ignored in Oswald et al.'s (2013) meta-analysis.

The virtue of Oswald and colleagues' approach is that it avoids cherry-picking findings that support a particular conclusion.[7] It also confirms that the predictive power of implicit (and explicit) measures is relatively small when person-, context-, and behavior-specific variables are ignored. Our point is that *this is to be expected*. In fact, as we stressed above, it would be bizarre to think otherwise, that is, to think that a generic measure of racial attitudes like the IAT would predict a wide range of race-related behavior irrespective of whether such a relation can be expected on theoretical grounds.

When the variables specified by theoretical models of implicit cognition are considered, it is clear that implicit measures are scientifically worthwhile instruments. For example, Cameron and colleagues (2012) analyzed 167 studies that used sequential priming measures. They found a small average correlation between sequential priming tasks and behavior (r = .28). Yet, much as Friese and colleagues (2008) predicted, correlations were substantially higher under theoretically expected conditions and close to zero under conditions where no relation would be expected. Cameron and colleagues identified their moderators from the fundaments of three influential dual-process models of social cognition.[8] 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.[9]

---

[6] For discussion of stereotype-attitude relations like these, see Fiske et al. (2002) and, in the context of implicit measures, Madva and Brownstein (2016).

[7] In another sense, though, Oswald and colleagues' focus on race and ethnicity, to the exclusion of all the other domains in which implicit measures have been found to predict behavior, *is* an instance of cherry-picking. They explicitly justify their focus on race and ethnicity IATs in terms of the popularity and media attention given to these measures, rather than on empirical grounds. And they do not acknowledge *a priori* reasons—that race is a socially sensitive topic and that "race-related" behavior is notoriously hard to operationalize—to expect lower predictive power from race and ethnicity-related IATs compared with other IATs.

[8] Specifically, from MODE model ("Motivation and Opportunity as Determinants;" Fazio 1990), APE model ("Associative-Propositional Evaluation;" Gawronski and Bodenhausen 2006), and MCM ("Meta-Cognitive Model;" Petty, Briñol, and DeMarree 2007).

[9] These points also apply to Forscher, Lai, and colleagues' (ms) meta-analysis of change in implicit measures. Their most publicized finding was that interventions that change performance on implicit measures like the IAT do not appear to lead to changes in behavior. Once again, however, Forscher, Lai, and colleagues did not code for theoretically important moderators. Without considering the predictions made by theoretical models about the conditions under which changes on implicit measures should and should not produce corresponding changes in behavior, one should *expect* to find little evidence of changes in the former being associated with changes in the latter. We also note that only 15% of the studies

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 and colleagues (ms) find an average correspondence of r = .37 in studies using the most effective IAT designs. Specifically, they find an average correspondence of r = .37 when they restrict their analysis to studies using a standard IAT rather than an IAT-variant, like the "Single-Category" IAT (Karpinski & 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 above, viz. recycling attitudes and recycling behavior). We discuss these findings, as well as additional ways to improve implicit measures, in §6. The point here is that, as follows from Cameron and colleagues' review, it is flatly unpersuasive to conclude anything from average zero-order correlations between measures of attitudes and behaviors, which ignores any theoretical expectations of the relations between them.

This is all the more the case when the incremental validity of implicit measures is taken into consideration. That is, these meta-analyses find that the IAT predicts behaviors over and above self-report measures (e.g., Greenwald et al. 2009a). Moreover, some specific studies that find no predictive power in self-report measures find significant predictive power in corresponding implicit measures (e.g., Agerström & Rooth 2011). Kurdi and colleagues (ms) replicated this result using a structural equation 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. Of course, our claim is *not* that implicit measures are generally better (or worse) than explicit measures. What remains the key open question is when—in what domains, under what conditions, etc.—implicit measures outperform explicit measures and when explicit measures outperform implicit measures. Each makes distinctive contributions to the prediction of behavior. Moreover, the conditions under which one type of measure outperforms the other will vary on theoretically expected grounds (e.g., when the topic is socially sensitive, when the motivation or opportunity to control spontaneous impulses is low, etc.).

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 (in press) 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 fit the data better than a single-construct model.[10] Indeed, even in the case of political attitudes, for which self-report measures are

included in Forscher, Lai, et al. (ms) included a behavioral outcome measure (and these included measures *intentions* to behave in such-and-such a way and measures that simply asked people how they *would* behave in a hypothetical situation) and only 22 studies measured attitude change over time. The dearth of relevant studies recommends hesitation in drawing strong conclusions.

[10] 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,

strongly predictive of political behavior, implicit measures have incremental validity. Friese and colleagues (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 single-target IAT showed incremental predictive validity. Greenwald and colleagues (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 attitudes (see Galdi et al. 2008).

4. Temporal Stability

Attitude-behavior relations should be small when all relevant effects sizes are averaged, irrespective of relevant theoretical and methodological considerations. This expectation is borne out in the above meta-analyses. This finding in no way impugns 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. But here, too, attention to various *a priori* and theoretically derived considerations, as well as the empirical evidence, clearly suggests that this challenge can be met with a more nuanced view.

4.1 Background

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.[11] 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. (See §5 for discussion of Payne and colleagues' "Bias of Crowds" model.) Critics of implicit bias research have taken the third route, arguing that test-retest instability suggests that measures like the IAT are faulty instruments.

The critics' case would be buoyed if implicit measures were unavoidably unstable, regardless of any other variables being held constant. Instability in measurement across time, under these

tapping into some single "implicit" construct (whether that construct is a process, representation, evaluation, etc.). See also Cunningham et al. (2001).

[11] However, it is important to note that changes across time on measures like the IAT may reflect relatively short term changes in the momentary accessibility of stored information or longer-term changes in the structure or strength of a person's associations themselves (Gawronski & Bodenhausen 2006, 2011; Madva 2016a).

conditions, would be indicative of a faulty tool. But this is not what the data suggest; implicit measures are not unavoidably unstable, as we discuss below.

4.2 Implicit Attitudes and Temporal Stability

In a recent longitudinal study, Gawronski and colleagues find 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 measures 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).[12] 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 merely 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 meaningful and temporally stable individual differences 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 which may differ from person to person, as well as over time, thereby producing low test-retest correlations. Gschwendner and colleagues (2008) illustrate 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

[12] Variance on any given measures can be divided into (1) systematic construct variance; (2) systematic measurement error; and (3) random error. Both construct variance and systematic measurement error contribute to internal consistency. Bar-Anan and Vianello (in press) 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 brings about a (non-random) change on an implicit measure, the effect could be related more to (2) than (1), i.e., changing the score without changing the construct of interest (by analogy, think of concerns in education about merely "teaching to the test"). This, together with the point we highlighted in the previous note, introduces the as-yet underexplored possibility that, for example, Forscher, Lai, et al.'s meta-analysis failed to find that changes in implicit bias correlated with changes in behavior because many of the relevant studies failed to meaningfully change implicit bias; changes in IAT scores might in many cases boil down to some combination of momentary changes in concept accessibility and systematic measurement error. This possibility is also frequently underappreciated in, e.g., debates and empirical research exploring whether automatic evaluations arise from associative versus propositional processes.

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 two-week period. Whereas participants in the control condition showed a relatively low stability coefficient of .29, participants in the experimental condition showed a relatively high stability coefficient of .72. This latter correlation is notably similar to Gawronski and colleagues' overall finding for stability of explicit attitudes (r = .75). This is one study, of course, and thus needs to be replicated and expanded upon. But it is highly suggestive that implicit measures are not unavoidably unstable. Rather, the conditions under which they are, and are not, stable must be better understood.

It may again bear emphasizing, however, that research along these lines long *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, activation, and priming. 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 (that is, the disposition to respond with certain thoughts, feelings, and behavioral impulses in a certain range of contexts). (See also discussion in §5.2 of Mischel & Shoda's (1995) comparable insights regarding personality research.)

There is additional suggestive evidence for relevant moderators of test-retest correlations elsewhere. Cooley and Payne (2016), 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 = .54 is an average correlation across all of the implicit measures they considered. For implicit political attitudes, however, they found a stability coefficient of .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 were decidedly lower—r = .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 are 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 (APE) Model (Gawronski & 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. More broadly, 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 into doubt.

5. Additional Worries

5.1 Hype

Critics have, sometimes rightly, pointed to the hype surrounded research on implicit bias as a cause for concern. We agree that research on implicit bias is often overhyped and mischaracterized, especially by the popular press[13], diversity consulting firms[14], and the websites of academic departments.[15] Such mischaracterizations can lead to the misuse of money intended to combat discrimination, the creation of misguided public policy, and popular misunderstanding of the workings of science.

Many (but not all) researchers have been more careful in describing the nature of the research. For example, legal scholars have been pushing researchers to investigate the use of the IAT in *voir dire*, but psychologists have pushed back, arguing that the measure is not appropriate for this application.[16] Likewise, Greenwald and colleagues (2015) caution against using the IAT as a diagnostic tool for classifying kinds of people (e.g., as "implicit racists"). In his "takedown" of implicit bias research, Jesse Singal characterizes this as a smoking gun, referencing the format in

---

[13] For example, Nicholas Kristof writes, "It's sobering to discover that whatever you believe intellectually, you're 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.

[14] For example, 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" to the problems associated with implicit bias. Indeed, this whole way of disentangling the two is misguided. See §5.4.

[15] For example, the Diversity and Cultural Competence website of the Johns Hopkins Medical School asserts that, "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.

[16] Banaji, p.c.

which individuals' results are given on the Project Implicit website after people take the IAT. Test-takers are told that their scores indicate "slight," "moderate," or "strong" bias.

While Project Implicit might do well to revise this form of feedback, it is not all the website says. In the FAQ section of Project Implicit, under the question, "What does it mean that my IAT score is labeled 'slight,' 'moderate,' or 'strong?'" the answer given reads, "If you respond faster when Flowers + Good/Insects + Bad are paired together compared to when Insects + Good/Flowers + Bad are paired together, we would say that you have an implicit preference for flowers relative to insects. The labels slight, moderate and strong reflect the strength of the implicit preference—how much faster you respond to Flowers + Good/Insects + Bad versus Insects + Good/Flowers + Bad." While we think it could be clearer, we take the point of this explanation to be that IAT feedback is not meant to diagnose individuals' deepest or most genuine attitudes, but rather, to explain that the labels are simply representations of error rates and latencies on the IAT itself.[17]

The broader point about the hype surrounding the IAT is that, from a psychometric perspective, it has no relevance for the quality of the research. Critics cannot impugn the research itself by pointing to the ways in which journalists, corporations, and politicians have misunderstood it, any more than they can 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] While responsible researchers have generally expressed sober caution about interpreting implicit bias, we grant that there have been instances of oversimplification and exaggeration. These instances must, however, be contextualized. First, they often represent misfires in the midst of good-faith efforts to make the complex psychometric and theoretical issues surrounding implicit social cognition more accessible to a wider audience. Second, 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

[17] Anecdotally, people often report feeling surprised by their IAT results. The long-standing interpretation of this has been that the IAT measures unconscious mental states. Hahn and colleagues (2014), however, found that naïve participants were quite accurate in predicting their implicit biases, and they cite a significant and expanding body of research suggesting that implicit biases are no less accessible to consciousness than explicit biases. Hahn and colleagues' data suggests that the feeling of surprise may be a product of the difference between participants' own subjective way of "measuring" their biases and the metrics used by the IAT. What many participants consider to be a "slight" bias, for example, is described as a "moderate" or "strong" bias on the Project Implicit website.

[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/

to these questions.[21] In any case, these communicative shortcomings have nothing to do with the scientific legitimacy of implicit measures.

5.2 Situationism and the Bias of Crowds

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 are very robust and stable; (2) children's average scores on 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 §§2-4, (4) individual differences in implicit bias have at best 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 implicit racial attitudes for each of the U.S. states, finding that, from one year to the next, the test-retest stability is quite high ($r = .76$), and remains so even over a ten-year span ($r = .69$). 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 and colleagues (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 (see Payne et al. (2017) for discussion)—underscore the need for careful study of the relations between implicit bias and social situations and structures.

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- and group-level data by appealing to the mercurial 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

[21] These glossing 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, etc. Given this, the temptation to describe implicit bias as "unconscious," for example, is understandable.

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," they write. "Instead, most of the systematic variance in implicit biases appears to operate at the level of situations."

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.[22] In seeing why, a comparison to research in personality psychology is instructive. Despite the binary uptake in recent philosophical discussion, which pits "persons" vs. "situations" (e.g., Harman 1999), it is a defining assumption of foundational theories that personality only emerges *in interaction with* situational variables (e.g., Lewin 1936; Bandura 1978; Mischel & 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). 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.[23]

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 vs. 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 & Banaji 2017). Population-level research, like Hehman and colleagues', treats

---

[22] Payne and colleagues themselves do not signal the death of the individual-differences approach, although some commenters attribute this view to them.

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

individual differences and short-term temporal instability in measurements as error. Here the object of study *is* the aggregate itself. 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, there are familiar risks associated with HARKing and p-hacking when statistical analyses become increasingly complex in these ways. 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, introduce more stringent tests of statistical significance, etc.). 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 et al. 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, 2017). 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) since none of these measures is "process-pure," different measures "tap into" different processes in theoretically expected ways.[24] Neither of these points undermines research on implicit bias.

Several reviews have found that not all indirect measures are equally reliable (Bar-Anan and Nosek 2014; Payne and Lundberg 2014). In general, the IAT and the AMP tend to do best in terms of their psychometric properties. 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—don't always strongly correlate. One explanation for this is that weak correlations reflect differences in 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.[25] 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 & 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 in press; cf. Payne and Lundberg 2014). Ultimately, the relatively low correlations between implicit measures is not so much an anomaly that threatens the field as it is a pedestrian empirical finding, which has begun to be explored (e.g., Moran et al. 2017, Bar-Anan and Nosek 2017). We would certainly welcome further theory-based and experimental investigation into the mechanisms explaining performance on these distinct measures (cf. Conrey et al. 2005; Bishara and Payne 2009; 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.

---

[24] There is also evidence that much of the *apparent* variance between implicit measures is simply another manifestation of the contextual variability of concept accessibility, i.e., the same phenomenon that explains the low test-retest reliability *within* implicit measures. For example, Cunningham and colleagues (2001) found much higher correlations between the IAT and the evaluative priming task after controlling for measurement error.

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

6. Future Directions

By no means is our claim that implicit measures are perfect. There is significant room for improvement. 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; Kurdi et al. 2017).[26] For example, Greenwald and colleagues (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 and colleagues' (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 and colleagues' (2009)] 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 .236 (or even .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 and colleagues (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%

[26] 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>.

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 measures of implicit attitudes 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 above)—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.[27]

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 above, Kurdi and colleagues (2017) 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 (2016) 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 (2016) 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

---

[27] 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 unmoving, given both that Greenwald and colleagues present the example explicitly as hypothetical and that 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 & 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, for example, (1991; Rosenthal and Rubin, 1982) shows that an *r* of .32 for a cancer treatment, compared to placebo, which accounts for only 10% of variance, translates into a survival rate of 66% in the treatment group compared to 34% in the placebo group.

This said, we note another caveat about the "accumulation mechanisms" defense of implicit bias research. Greenwald et al.'s (2015) statistical model is based on the assumption of additivity, and there are good reasons to assume a multiplicative instead of an additive model. In a multiplicative model, "trickle down effects" actually 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.

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 comparatively 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 arguably successful by targeting specific, contextually relevant associations in theoretically informed ways.

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 (Tetlock et al. 2000; Jost 2017), on "fast and frugal" heuristics and biases (Hewstone, Benn, and Wilson 1988; Kahneman 2011; Gigerenzer 2008; Peer and Gamliel 2013), on the dependency relations in networks of concepts (Sloman, Love, and Ahn 1998; Meyer et al. 2015; del Pinal, Madva, and Reuter 2017), 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 will surely be relevant to explaining and predicting contemporary prejudice and discrimination, but many of them may 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 (Holland et al. 2012, Forgas 2011). More research on moderators like these is needed (cf. Madva 2017).

Research on implicit bias has certainly been overhyped by some. Worse, some have taken the IAT to be a diagnostic tool for classifying kinds of people (e.g., as "implicit racists"). But these mistakes do not impugn the scientific validity of the underlying research. Such research should proceed and continue to be taken seriously by philosophers, psychologists, and the public.

Works Cited

Agerström, J., and Rooth, D. O. (2011). The role of automatic obesity stereotypes in real hiring discrimination. *Journal of Applied Psychology*, *96*, 790–805.

Ajzen, I., & Fishbein, M. (1977). Attitude-behavior relations: A theoretical analysis and review of empirical research. *Psychological Bulletin, 84,* 888-918.

Amodio, D. M., & Devine, P. G. (2006). Stereotyping and evaluation in implicit race bias: evidence for independent constructs and unique effects on behavior. *Journal of Personality and Social Psychology*, *91*(4), 652.

Axt, J.R. Manuscript. "The Best Way to Measure Explicit Racial Attitudes Is to Ask about Them." https://osf.io/se9tx/.

Banks, R. R., & Ford, R. T. (2008). (How) Does Unconscious Bias Matter: law, Politics, and Racial Inequality. *Emory LJ*, *58*, 1053.

Bandura, A. (1978). The self system in reciprocal determinism. *American Psychologist, 33,* 344-358.

Bar-Anan, Y., & Nosek, B. 2014. A comparative investigation of seven implicit attitude measures. *Behavioral Research*, 46, 668-688.

Bar-Anan, Y., & Nosek, B. A. (2017). A Comparison of the Sensitivity of Four Indirect Evaluation Measures to Evaluative Information.

Bar-Anan, Y., & Vianello, M. (in press). A Multi-Method Multi-Trait Test of the Dual-Attitude Perspective. Retrieved from *PsyArXiv*, https://doi.org/10.17605/OSF.IO/BJAQ8

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., & Payne, B. K. (2009). Multinomial process tree models of control and automaticity in weapon misidentification. *Journal of Experimental Social Psychology*, *45*(3), 524–534.

Bodenhausen, G. & Gawronski, B. 2017. Beyond Persons and Situations: An Interactionist Approach to Understanding 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., & Madva, A. 2012. The normativity of automaticity. *Mind and Language*, 27(4), 410-434.

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

Cervone, D., Shadel, W., and Jencius, S. 2001. Social-Cognitive Theory of Personality Assessment. *Personality and Social Psychology Review* 5:1, 33-51.

Cesario, Joseph, and Kai J. Jonas. 2014. "Replicability and Models of Priming: What a Resource Computation Framework Can Tell Us About Expectations of Replicability." Social Cognition 32 (Supplement): 124–36. doi:10.1521/soco.2014.32.supp.124.

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. *Psychological Science* 21:9, 1311-1317.

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), p. 70.

Conrey, F., J. Sherman, B. Gawronski, K. Hugenberg, & C. Groom, 2005, "Separating multiple processes in implicit social cognition: The Quad-Model of implicit task performance", *Journal of Personality and Social Psychology*, 89: 469–487.

Cooley, E., and B. K. Payne. 2016. "Using Groups to Measure Intergroup Prejudice." *Personality and Social Psychology Bulletin*, November. doi:10.1177/0146167216675331.

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

del Pinal, Guillermo, Alex Madva, and Kevin Reuter. 2017. "Stereotypes, Conceptual Centrality and Gender Bias: An Empirical Investigation: Stereotypes, Conceptual Centrality and Gender Bias." Ratio, June. doi:10.1111/rati.12170.

Devine, P., Forscher, P., Austin, A, Cox, W. 2012. Long-term reduction in implicit race bias; a prejudice habit-breaking intervention. *Journal of Experimental Social Psychology,* 48(6), 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, 411-425.

Fazio, R. H. (1990). Multiple processes by which attitudes guide behavior: The MODE model as an integrative framework. *Advances in Experimental Social Psychology, 23,* 75-109.

Fazio, R. & Williams, C. 1986. Attitude Accessibility as a Moderator of the Attitude-Perception and Attitude-Behavior Relations: An Investigation of the 1984 Presidential Election. *Journal of Personality and Social Psychology* 51:3, 505-514.

Fleeson, W. 2004. Moving Personality Beyond the Person-Situation Debate. *Current Directions in Psychological Science* 13:2, 83-87.

Fiske, S. T., Cuddy, A. J., Glick, P., and Xu, J. (2002). A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. *Journal of Personality and Social Psychology*, *82*, 878.

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. Manuscript. A Meta-Analysis of Change in Implicit Bias.

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

Friese, M., Hofmann, W., & Schmitt, M. (2008). When and why do implicit measures predict behaviour? Empirical evidence for the moderating role of opportunity, motivation, and process reliance. *European Review of Social Psychology*, *19*(1), 285–338. https://doi.org/10.1080/10463280802556958

Galdi, S., Arcuri, L., & Gawronski, B. (2008). Automatic mental associations predict future choices of undecided decision-makers. Science, 321, 1100-1102.

Gawronski, B., & Bodenhausen, G. V. (2006). Associative and propositional processes in evaluation: An integrative review of implicit and explicit attitude change. *Psychological Bulletin, 132,* 692-731.

Gawronski, B., & Bodenhausen, G.V. 2011. The associative-propositional evaluation model. Theory, evidence, and open questions. *Advances in Experimental Social Psychology*, 44, 59-127.

Gawronski, B., & Cesario, J. 2013. Of mice and men: What animal research can tell us about context effects on automatic response in humans. *Personality and Social Psychology Review*, 17(2), 187-215.

Gawronski, B., & De Houwer, J. (2014). Implicit measures in social and personality psychology. In H. T. Reis, & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (2nd edition, pp. 283-310). New York, NY: Cambridge University Press.

Gawronski, B., Morrison, M., Phills, C., and Galdi, S. 2017. Temporal Stability of Implicit and Explicit Measures: A Longitudinal Analysis. *Personality and Social Psychology Bulletin* 43:3, 300-312.

Gawronski, B., & Strack, F. (Eds.). (2012). *Cognitive Consistency: A Fundamental Principle in Social Cognition* (1 edition). New York, NY: The Guilford Press.

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

Greenwald, A. G., Banaji, M. R., & Nosek, B. A. (2015). Statistically Small Effects of the Implicit Association Test Can Have Societally Large Effects. *Journal of Personality and Social Psychology, 108,* 553–561.

Greenwald, A., McGhee, D., & Schwartz, J. 1998. Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology*, 74, 1464-1480.

Greenwald, A., Poehlman, T., Uhlmann, E., & M. Banaji. 2009a. Understanding and using the implicit association test: III meta-analysis of predictive validity. *Journal of Personality and Social Psychology*, 97(1), 17-41.

Greenwald, A. G., Poehlman, T., Luis Uhlmann, E., & Banaji, M. (2009). Understanding and Using the Implicit Association Test: III. Meta-Analysis of Predictive Validity, *97*(1), 17–41. https://doi.org/10.1037/a0015575

Gschwendner, T., Hofmann, W., & 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,* 70-79.

Hahn, A., Judd, C. M., Hirsh, H. K., & Blair, I. V. (2014). Awareness of implicit attitudes. *Journal of Experimental Psychology: General*, *143*(3), 1369.

Harman, G. 1999. Moral philosophy meets social psychology: Virtue ethics and the fundamental attribution error. *Proceedings of the Aristotelian Society*, 99, 315–331.

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

Hehman, E., Flake, J. K., & Calanchini, J. (2017). Disproportionate use of lethal force in policing is associated with regional racial biases of residents. *Social Psychological and Personality Science*, 1948550617711229.

Hewstone, M., Benn, W., & Wilson, A. (1988). Bias in the use of base rates: Racial prejudice in decision-making. *European Journal of Social Psychology*, *18*(2), 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* (1 edition). Farrar, Straus and Giroux.Karpinski, A., & Steinman, R. B. (2006). The single category implicit association test as a measure of implicit social cognition. *Journal of personality and social psychology*, *91*(1), 16.

Kurdi, B. & Banaji, M. 2017. Reports of the death of the individual difference approach to implicit social cognition may be greatly exaggerated: A commentary on Payne, Vuletich, & Lundberg (2017).

Kurdi, B., Seitchik, A., Axt, J., Carroll, T., Karapetyan, A., Kaushik, N., Tomezsko, D., Greenwald, A., and Banaji, M. Manuscript. Relationship between the Implicit Association Test and intergroup behavior: A meta-analysis.

Leslie, S.-J., Cimpian, A., Meyer, M., & Freeland, E. (2015). Expectations of brilliance underlie gender distributions across academic disciplines. *Science*, *347*(6219), 262–265. https://doi.org/10.1126/science.1261375

Levinson, J. D., Smith, R. J., & 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*, 513-581.

Lewin, K. (1936). A DYNAMIC THEORY OF PERSONALITY: SELECTED PAPERS. *The Journal of Nervous and Mental Disease*, *84*(5), 612-613.

Loersch, Chris, and B. Keith Payne. 2011. "The Situated Inference Model: An Integrative Account of the Effects of Primes on Perception, Behavior, and Motivation." Perspectives on Psychological Science 6 (3): 234–52. doi: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

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

Machery, E. 2016. De-Freuding Implicit Attitudes. In M. Brownstein and J. Saul (Eds.), *Implicit Bias and Philosophy: Volume 1, Metaphysics and Epistemology*, 104-129. Oxford, UK: Oxford University Press.

Machery, E. 2017. Do Indirect Measures of Bias Measure Traits or Situations?

Madva, A. (2016a). Why implicit attitudes are (probably) not beliefs. *Synthese*, *193*(8), 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), 701–728. https://doi.org/10.3998/ergo.12405314.0003.027

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

Madva, A. 2017. Implicit Bias, Moods, and Moral Responsibility. *Pacific Philosophical Quarterly*. DOI: 10.1111/papq.12212

Madva, A. and Brownstein, M. 2016. Stereotypes, Prejudice, and the Taxonomy of the Implicit Social Mind. *Noûs*. DOI: 10.1111/nous.12182.

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

Mekawi, Y., & Bresin, K. 2015. Is the evidence from racial bias shooting task studies a smoking gun? Results from a meta-analysis. *Journal of Experimental Social Psychology*, *61*, 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, 41–749.

Meyer, M., Cimpian, A., & 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., & Shoda, Y. (1995). A cognitive-affective system theory of personality: reconceptualizing situations, dispositions, dynamics, and invariance in personality structure. *Psychological review*, *102*(2), 246.

Moran, Tal, Yoav Bar-Anan, and Brian A. Nosek. 2017. "The Effect of the Validity of Co-Occurrence on Automatic and Deliberate Evaluations." European Journal of Social Psychology, n/a-n/a. doi:10.1002/ejsp.2266.

Nosek, B. A. (2007). Implicit–Explicit Relations. *Current Directions in Psychological Science*, *16*(2), 65–69. https://doi.org/10.1111/j.1467-8721.2007.00477.x

Oskamp, S., Harrington, M. J., Edwards, T. C., Sherwood, D. L., Okuda, S. M., & Swanson, D. C. (1991). Factors influencing household recycling behavior. *Environment and behavior*, *23*(4), 494-519.

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

Payne, B. K., Hall, D. L., Cameron, C. D., & Bishara, A. J. (2010). A Process Model of Affect Misattribution. *Personality and Social Psychology Bulletin*, *36*(10), 1397–1408. https://doi.org/10.1177/0146167210383440

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

Payne, B.K., Cheng, C.M., Govorun, O., & Stewart, B.D. 2005. An inkblot for attitudes: Affect misattribution as implicit measurement. *Journal of Personality and Social Psychology*, 89(3), 277-293.

Payne, K., Vuletich, H., and Lundberg, K. 2017. Flipping the Script on Implicit Bias Research with the Bias of Crowds.

Peach, Jennifer M., Emiko Yoshida, Steven J. Spencer, Mark P. Zanna, and Jennifer R. Steele. 2011. "Recognizing Discrimination Explicitly While Denying It Implicitly: Implicit Social Identity Protection." Journal of Experimental Social Psychology 47 (2): 283–92. doi:10.1016/j.jesp.2010.09.007.

Peach, Jennifer M., Emiko Yoshida, and Mark P. Zanna. 2011. "Learning What Most People like: How Implicit Attitudes and Normative Evaluations Are Shaped by Motivation and Culture and Influence Meaningful Behavior." In The Psychology of Attitudes and Attitude Change, edited by Joseph P. Forgas, Joel Cooper, William D. Crano, and Steven J. Spencer, 95–108. Philadelphia: Psychology Press.

Peer, E., & Gamliel, E. (2013). Heuristics and Biases in Judicial Decisions. *Court Review*, *49*, 114.

Petty, R. E., Briñol, P., & DeMarree, K. G. (2007). The meta-cognitive model (MCM) of attitudes: Implications for attitude measurement, change, and strength. *Social Cognition, 25,* 657-686.

Poropat, A.E. (2009). A Meta-Analysis of the Five-Factor Model of Personality and Academic Performance. Psychological Bulletin, 135(2), 322-338. doi:10.1037/a0014996

Richardson, M., Abraham, C., & Bond, R. (2012). Psychological correlates of university students' academic performance: a systematic review and meta-analysis. Psychological Bulletin, 138, 353- 387. doi:10.1037/a0026838

Rentfrow, Peter J., Markus Jokela, and Michael E. Lamb. 2015. "Regional Personality Differences in Great Britain." *PLOS ONE* 10 (3): e0122245. doi: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. *Journal of Personality and Social Psychology* 90:1, 109-128.

Rosenthal, R. 1991. Meta-analytic procedures for social research (Rev. ed.). Newbury Park, CA: Sage.

Rosenthal, R., & Rubin, D.B. 1982. A simple, general purpose display of magnitude of experimental effect. *Journal of Educational Psychology*, 74, 166 –169.

Samayoa and Fazio (2017). Reply to Payne et al. Bias of Crowds.

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, Steven A., Bradley C. Love, and Woo-Kyoung Ahn. 1998. "Feature Centrality and Conceptual Coherence." Cognitive Science 22 (2): 189–228. doi:10.1207/s15516709cog2202_2.

Strack, F., & Deutsch, R. (2004). Reflective and impulsive determinants of social behavior. *Personality and Social Psychology Review*, *8*, 220-247.

Strenze, T. Intelligence and socioeconomic success: A meta-analytic review of longitudinal research. *Intelligence* 35, 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, 263-296.

Tetlock, P. E., Kristel, O. V., Elson, S. B., Green, M. C., & Lerner, J. S. (2000). The psychology of the unthinkable: Taboo trade-offs, forbidden base rates, and heretical counterfactuals. *Journal of Personality and Social Psychology*, *78*(5), 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, 198-213.

Walton, G.M., C. Logel, J.M. Peach, S.J. Spencer, and M.P. Zanna. 2015. "Two Brief Interventions to Mitigate a 'chilly Climate' Transform Women's Experience, Relationships, and Achievement in Engineering." Journal of Educational Psychology 107 (2): 468.Westfall, J., & Yarkoni, T. (2016). Statistically Controlling for Confounding Constructs Is Harder than You Think. *PLoS ONE*, *11*(3), e0152719–22. http://doi.org/10.1371/journal.pone.0152719

Wolfe, R., & Johnson, S., 1995. Personality as a Predictor of College Performance. *Education and Psychological Measurement* 55:2, 177-185.

Yao, V. & Reis-Dennis, S. 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, Emiko, Jennifer M. Peach, Mark P. Zanna, and Steven J. Spencer. 2012. "Not All Automatic Associations Are Created Equal: How Implicit Normative Evaluations Are Distinct from Implicit Attitudes and Uniquely Predict Meaningful Behavior." Journal of Experimental Social Psychology 48 (3): 694–706. doi:10.1016/j.jesp.2011.09.013.

Zanna, M. P., & Fazio, R. H. (1982). The attitude-behavior relation: Moving toward a third generation of research. In M. P. Zanna, E. T. Higgins, C. P. Herman (Eds.), *Consistency in social behavior: The Ontario symposium* (Vol. 2, pp. 283-301). Hillsdale, N.J.: Erlbaum.