What Do Implicit Measures Measure?

Michael Brownstein
John Jay College/CUNY

Alex Madva
Cal Poly Pomona

Bertram Gawronski
University of Texas at Austin

We identify several ongoing debates related to implicit measures, surveying prominent views and considerations in each debate. First, we summarize the debate regarding whether performance on implicit measures is explained by conscious or unconscious representations. Second, we discuss the cognitive structure of the operative constructs: are they associatively or propositionally structured? Third, we review debates whether performance on implicit measures reflects traits or states. Fourth, we discuss the question of whether a person's performance on an implicit measure reflects characteristics of the person who is taking the test or characteristics of the situation in which the person is taking the test. Finally, we survey the debate about the relationship between implicit measures and (other kinds of) behavior.

Research on implicit social cognition suggests that people can act in biased ways without intending to do so. The advent of implicit measures has given researchers tools to capture unintended biases in precise and empirically tractable ways. Most generally, implicit measures assess people's thoughts and feelings without having to ask them directly, "what do you think about X?" The most well-known implicit measure—the Implicit Association Test (IAT; Greenwald et al. 1998)—accomplishes this by measuring the speed and accuracy with which people sort different kinds of stimuli (e.g., images of black and white men and positive and negative words) into different categories (e.g., *European American*, *African American*, *Good*, and *Bad*). When participants are forced to sort target objects in ways that conflict with common social stereotypes and prejudices—for example, by responding to "good" words and black faces with the same key—they tend to show slower responses and make more mistakes. Other implicit measures, such as the Affect Misattribution Procedure (AMP; Payne et al. 2005), ask participants to make evaluative judgments about ambiguous target images (e.g., Chinese pictographs for English speakers) after priming them with briefly presented stimuli (e.g., images of black men). Extensive research suggests that people's evaluations of the ambiguous target stimuli are systematically affected by the presentation of primes, despite being told to ignore them.

These tools have been explored, deployed, theorized about, and recently, criticized extensively. One set of critiques has focused on psychometric properties of the IAT, such as how scores are computed, whether there are precise cutoff points between being "biased" and "unbiased," and so on (e.g., Blanton & Jaccard 2006; Mitchell & Tetlock 2017). Another set of critiques has focused on more widely-applicable criteria of psychometric strength, such as how well the IAT and other implicit measures predict behavior, whether they have acceptable test-retest reliability, and whether changes in implicit bias cause changes in behavior (e.g., Forscher, Lai et al. ms; Machery 2016, 2017; Oswald et al. 2013). At the same time, some philosophers, political theorists, and other social scientists have developed various "structuralist" critiques of implicit bias research (e.g., Anderson 2010; Ayala 2016; 2018; Dixon et al. 2012; Dixon & Levine 2012; Haslanger 2015; Mallon 2018). The core idea here is that disparate social outcomes and ongoing intergroup discrimination—the phenomena implicit bias research putatively aims to explain—are caused by structural features of society, not by biased individual minds. In fighting for equity, so the critique goes, our focus ought to be on poverty, housing segregation, social norms, mass incarceration, etc. Other critics have argued that implicit bias is unimportant compared with explicit bias (Hermanson 2017) and that the implicit bias research program has been overhyped (Blanton & Ikizer 2018; Jussim 2018).

We are more sympathetic to some of these criticisms than others. Elsewhere we have addressed specific lines of criticism in detail (Brownstein forthcoming; Brownstein et al. ms; Gawronski, forthcoming; Madva 2016a, 2017). Here we step back from any one set of concerns in order to try to reframe the discussion a bit. We ask: *what do implicit measures measure*? By far, the most common way to answer this question is by postulating a specific kind of mental construct, one which putatively defines the nature of implicit bias. According to Greenwald and Banaji's (1995) influential definition, for example, implicit bias reflects "introspectively unidentified (or inaccurately identified) traces of past experience that mediate favorable or unfavorable feeling, thought, or action toward social objects" (p. 8). Other theorists have proposed alternatives. Proposals include claims that implicit biases should be understood as conscious propositional representations (De Houwer 2014), unconscious beliefs (Mandelbaum 2016), attitudes that affect behavior under certain conditions (Olson & Fazio 2008; Petty et al. 2007), previously learned attitudes that co-exist with newly formed attitudes (Wilson et al. 2000), and arational mental states with representational, affective, and behavioral components (Brownstein & Madva 2012; Brownstein 2018; Gendler

2008a,b; Madva & Brownstein 2018). To some extent, the existence and influence of these competing definitions are the product of the multiple historical roots of research on implicit social cognition (e.g., research on implicit memory, on the one hand, and research on automatic processes in attention, on the other; for discussion of the relevant history, see Greenwald & Banaji 2017; Payne & Gawronski 2010). All of these proposals presume that implicit measures capture implicit bias and that implicit bias is some kind of mental construct. Another way to put this is that there is a general presumption in the literature that implicit measures are proxies for (contestably defined) mental constructs. Much of the criticism has thereby focused on whether these kinds of mental constructs exist, whether they can be adequately measured, and whether they have real-world significance. Do the data support the claim that people have unconscious attitudes? Do unconscious attitudes predict real-world behavior? And so on.

Rather than treat implicit measures as capturing proxies for mental states, we begin with the observation that implicit measures assess *behavior*. The IAT measures the speed and accuracy with which people sort pictures and words on a computer screen by pressing buttons on a keyboard; the Evaluative Priming Task (EPT; Fazio et al., 1995) measures the speed and accuracy with which people classify positive and negative target words following the presentation of primes; and the AMP measures participants' evaluations of ambiguous stimuli following the presentation of primes. Keeping in mind that implicit measures capture various kinds of behavior allows us to treat statements about mental constructs as theoretical hypotheses that seek to explain test performance rather than as objects of assessment themselves. We propose that the status of implicit measures can best be assessed from this perspective. In other words, rather than begin with a theory of the psychological nature of implicit bias and then assess the strength of tests of this construct and their relationship to ecologically significant criteria (e.g., behavior outside the lab), we will consider the "behavioral" data itself (i.e., performance on categorization tasks like the IAT), what sorts of *other* outcomes it does and does not predict, and what sorts of hypotheses about the underlying psychological architecture these relationships do and do not support.[1]

Here is an analogy to clarify our approach: someone might hypothesize that performance in the decathlon predicts athletic performance in other sports. This theorist need not, at the start, identify what the decathlon is actually measuring "underneath" the athletic performance itself. The question is whether one set of athletic performances predicts other sets of athletic performances. The question of what specifically the decathlon is measuring is simply *another* question (or set of questions), constrained but not determined by research correlating decathlon performance with other behaviors. One of our guiding aims in this article is to emphasize that these are two distinct—but mutually informing—sets of empirical projects: (1) how different types of behavioral performance correlate with each other and (2) what might causally underlie either of those performances and their relation.

In this article, we identify several debates related to implicit measures, surveying prominent views and considerations in each debate. First, we summarize the debate regarding whether performance on implicit measures is explained by conscious or unconscious representations. Second, we discuss the cognitive structure of the operative constructs: are they associatively or propositionally structured? Third, we review debates whether performance on implicit measures reflects traits or states. Fourth, we discuss the question of whether a person's performance on an implicit measure reflects characteristics of the person who is taking the test or characteristics of the situation in which the person is taking the test. Finally, we survey the debate about the relationship between implicit measures and (other kinds of) behavior.

Conscious or Unconscious Representations

Some influential theories define implicit bias in terms of unconscious attitudes. This characterization is the one largely used in trade books, newspapers, magazines, and so on, where the impression is often given that implicit bias is synonymous with "unconscious bias." Greenwald and Banaji's (1995) definition (quoted above), which defines implicit bias in terms of introspective failures to identify the elements of past experiences that influence one's current feelings, thoughts, and behavior, has been highly influential in popularizing this idea. More recently, however, Greenwald and Banaji (2017) adopted an approach closer to the one we recommend here. They argue that "implicit" ought to be used in an "empirical" rather than "conceptual" sense. The empirical sense refers to studies employing indirect techniques for measurement (i.e., without directly asking subjects what they think about some topic). The conceptual sense refers to a certain kind of mental construct (viz. an unconscious one).

There are two principal reasons to support the hypothesis that what explains performance on measures like the IAT are introspectively inaccessible (or inaccessed) representations. The first derives from anecdotal reports that when people are informed of their results on these tests, they often experience or express surprise and consternation (e.g., Banaji, 2011; Krickel,

[1] For discussion of related proposals, see De Houwer et al. (2013) and Johnson (ms).

2018). Such anecdotes have been taken to suggest that one *discovers* one's implicit biases (viz. unconscious attitudes) by taking the relevant tests, just as one might discover that one has high cholesterol by having a blood test. The second reason is that meta-analyses have found relatively low correlations between implicit and explicit measures (r = .2-.25; Cameron et al. 2012; Hofmann et al. 2005). This finding suggests significant dissociation between people's putative implicit and explicit biases. The hypothesis that people are unaware of their implicit biases could explain this dissociation.

However, conceptual and empirical reasons have been given for questioning whether implicit biases are unconscious. One conceptual concern is that it is unclear what exactly people fail to introspect. As Gawronski and colleagues (2006) pointed out, people may be unaware of the source of their attitudes (e.g., the past experiences that shape their evaluations of black and white people), the content of their attitudes (e.g., what their evaluations of black and white people are), or the effects their attitudes have on their behavior (e.g., how they evaluate CVs of equally qualified black and white candidates). Yet, even for openly reported preferences, such as a love of coffee or a hatred of cilantro, people might not know all their sources and effects. An avowed racist who openly disparages nonwhite people might have no idea where his views came from, and he might sincerely deny that his negative treatment toward a *particular* black individual (such as a job applicant) was influenced by his general negative attitudes. Thus, with regard to the distinction between implicit and explicit measures, the critical question is whether people are unaware of the content of their attitudes, as they are reflected on implicit measures.

The reasons given for lack of content awareness are themselves open to interpretation. For example, the experience of surprise after being told that one has certain implicit biases may be an artifact of the form in which test feedback is given (Gawronski, forthcoming; Hahn et al., 2014). It is also unclear to what extent people experience surprise or are merely expressing surprise given perceived social pressure to do so, or simply experience defensiveness because being accused of racial biases that run counter to one's explicit self-conception is painful (e.g., Howell, Gaither, and Ratliff 2015; Howell and Ratliff 2017). It is further unclear *how many* people experience surprise, as evidence for this is primarily anecdotal; it could be, for example, a matter of individual differences, such that some people are less disposed to introspective self-examination than others and so are more surprised to learn about biases that other people notice in themselves more regularly. Relatedly, perhaps most people *can* access their attitudes, but only some *do*.

There is also more to be said about meta-analyses demonstrating dissociations between implicit and explicit measures. One possibility is that these dissociations are explained by the principle of correspondence (Ajzen & Fishbein 1977). If the attitude-objects of explicit measures and the attitude-objects of implicit measures are not well matched, then the principle of correspondence predicts that they will not be strongly correlated. Existing meta-analyses that have coded for implicit/explicit correspondence have indeed found higher correlations between implicit and explicit measures when they matched in terms of the correspondence principle (Hofmann et al. 2005; see also Payne et al. 2008 for discussion of the influence of structural fit on correspondence between implicit and explicit measures). Finally, a series of experiments puts additional pressure on the hypothesis that unconscious attitudes explain performance on implicit measures. When people are told that the IAT is "as close to a lie detector test as is possible," for example, there is significantly stronger association between their IAT scores and their scores on corresponding explicit measures (Nier 2005). Recent studies also found that simply telling participants to attend to their "gut feelings" or spontaneous affective reactions toward minority groups significantly increased their acknowledgment of bias (Cooley et al. 2015; Lee, Lindquist & Payne 2017), and to just the same extent as asking participants to predict their IAT scores—even if participants did not actually take IATs (Hahn et al. 2014; Hahn & Gawronski 2018; Rivers & Hahn 2018). Finally, research by Hahn et al. (2014) suggests that people are surprisingly accurate in predicting their scores on future IATs, which poses a challenge to the hypothesis that people are unaware of their implicit biases.

As evidence for people's awareness of their implicit biases accumulates, "hard evidence that people have attitudes and beliefs that they don't know about, or can't know about when they try, is difficult to find" (Hall & Payne 2010). Theorists of course do not deny that people have unconscious biases or that a great proportion of mental life is consciously inaccessible. The debate here is whether implicit measures per se assess such unconscious processes or representations. If consensus continues to grow that implicit biases are consciously accessible, it is fair to ask what makes them "implicit" at all. Diversity trainers and journalists often refer to "implicit bias" and "unconscious bias" interchangeably. However, the explicit/implicit divide in cognition has historically been associated with a variety of other contrasts, including controlled versus automatic, slow versus fast, associative versus non-associative, effortful versus effortless, and voluntary versus involuntary. (See Brownstein (2018) for historical and contemporary usage of the term "implicit" in philosophy and psychology.) Moreover, as we explained in the introduction, a more minimal and less theoretically committal interpretation is available, wherein "implicit" simply refers to tests that do not explicitly ask participants to report their thoughts and feelings (as in, "implicit measures").

Associative and/or Propositional Representations

Perhaps equally influential as the idea that implicit measures capture unconscious representations is the idea that they capture associations. These accounts are orthogonal, as associations could be either conscious or unconscious. The hypothesis that associations of some kind (whether conscious or unconscious) are responsible for test performance is pervasive in the literature. As Kelly and Roedder (2008) and Mandelbaum (2016) have pointed out, the IAT stands for the Implicit *Association* Test, after all. Leading theorists repeatedly define "implicit bias" in terms of associative mental states. Nosek and colleagues (2007), for example, write that "the IAT (Greenwald et al. 1998) assesses associations between two concepts (e.g., black people and white people) two attributes (e.g., good and bad)" (8; although note that elsewhere Nosek has been more agnostic about the structure of implicit representations (e.g., Greenwald et al. 2005)).

One deflationary interpretation is that "associative" and its cognates are used in the literature as theoretically noncommittal placeholders for whatever structure the mental constructs assessed by implicit measures have (Greenwald & Banaji 2017). But the more pervasive interpretation situates "associations" as part of dual-system and/or dual-process psychology. In early social cognition research, associations were taken to be characterized by slow-learning and relative inflexibility (e.g., Smith & DeCoster 2000). More recent theorizing (e.g., Strack & Deutsch's 2004 Reflective-Impulsive Model (RIM)) interprets performance on implicit measures in terms of an "impulsive system" that influences approach and avoidance behaviors by "spreading activation" within associative networks. This spreading-activation mechanism contrasts with the mechanism underlying a "reflective system" that influences behavior through propositional reasoning.

Gawronski, Brannon, and Bodenhausen (2017) recommend distinguishing three senses in which associative constructs might explain test performance. (See De Houwer 2009 and Mandelbaum 2016 for related discussion.) First, the processes by which an agent learns information may be associative. The core idea of associative learning is that observed co-occurrences create direct links between concepts in memory. The alternative to this view is propositional learning, which is sensitive to the perceived truth or falsity of observed relations between co-occurring events (e.g., stimulus A starts vs. stops the co-occurring stimulus B). Second, the structure of representations stored in memory may be associative. The core idea of associative structures is that token associations are defined in terms of the relative strength of the links between nodes (e.g., that A is strongly or weakly linked to B). Propositional structures, by contrast, specify how concepts are related (e.g., that A *causes* B or that A *prevents* B). Representations with propositional structures can capture the truth or falsity of information, while representations with associative structure cannot (see discussion below). Relatedly, representations with propositional structure can be subject to inferential transitions between thoughts (e.g., modus tollens), while representations with associative structure cannot. Third, the processes by which representations influence behavior may be associative. In this sense, associative processes of feature-matching and spreading-activation have been assumed to underlie the activation of stored representations, while propositional processes are concerned with the validation of activated representations (Gawronski & Bodenhausen 2006). The core idea here is that associatively activated representations can directly influence spontaneous reactions to an object even when the content of these representations is rejected as invalid. In contrast, deliberate actions are assumed to be the product of reflectively generated intentions based on activated information that is deemed valid (Strack & Deutsch, 2004).

The key question for assessing whether there is evidence for an associative-propositional distinction in *learning* has to do with sensitivity or insensitivity to information about the substantive relationships between co-occurring stimuli. Specifically, do people link co-occurring stimuli regardless of the relationship between those stimuli? It is thought to be evidence of dissociation between associative and propositional processes in learning if people link co-occurring stimuli regardless of their relation. For example, Moran and Bar-Anan (2013) presented participants with a neutral stimulus that either started or stopped a pleasant or unpleasant sound. On an explicit measure, participants preferred stimuli that started pleasant sounds to stimuli that started unpleasant sounds, and conversely preferred stimuli that stopped unpleasant sounds to stimuli that stopped pleasant sounds. But on an implicit measure, participants' evaluations simply reflected the co-occurrence of stimuli; they preferred the stimuli that were paired with pleasant sounds to the stimuli that were paired with unpleasant sounds, regardless whether the stimuli started or stopped the sounds. While some studies have found similar dissociations (e.g., Hu et al. 2017, Experiments 1 and 2), others have failed to find it, showing effects of relational information on both explicit and implicit measures (e.g., Gawronski, Walther, & Blank, 2005; Hu et al. 2017, Experiment 3). At this point, the critical question is whether these inconsistent findings are better explained in terms of two functionally distinct learning mechanisms (i.e., associative vs. propositional learning) or rather by other factors, such as how stored relational information is retrieved, influencing how behavior is expressed (Hu et al., 2017; Van Dessel, Gawronski, & De Houwer, forthcoming).

Distinguishing between *representations* with associative and propositional structure depends on background theories about what kinds of mental systems can represent complex relations between stimuli, such as how the stimuli are related. De Houwer (2009) and Mandelbaum (2016) argue, for example, that associative structures cannot represent complex relations between stimuli, while Gawronski and Bodenhausen (2018) argue that they can. Here we note only the difficulty of assessing the structure of hypothesized mental representations from behavioral data (see De Houwer et al., 2013). For example, Mandelbaum (2016) argues that, "we can infer whether a given cognitive structure is associative by seeing how certain types of information modify (or fail to modify) behaviors under the control of the cognitive structures" (6). Many studies have proceeded in this vein, by exploring whether verbal information with propositional content (such as information about the consequences of a certain action) exerts an effect on performance on implicit measures (e.g., Mann & Ferguson 2017; Van Dessel, De Houwer, and Smith 2018). We recommend caution, however, in interpreting these kinds of findings as unambiguous evidence regarding the *structure* of mental representations. First, some leading theories that appeal to associative processes have from the outset proposed causal pathways by which propositional processes can influence associative processes, and vice versa (Gawronski and Bodenhausen 2006). Second—and this returns us to our central point of emphasis in this paper that implicit measures assess behavior—evidence that propositionally rich information changes performance on, say, an IAT or AMP tells us relatively little about the specific underlying constructs responsible for that behavioral performance. For example, with regard to the relative impact of co-occurrence and relational information, potential dissociations between implicit and explicit measures may reflect differences in (a) learning mechanisms, (b) mental representations, or (c) processes during the expression of a behavioral response (Hu et al., 2017). In line with this concern, Van Dessel et al. (forthcoming) pointed out that differential effects of co-occurrence and relational information on implicit and explicit measures may not be due to two functionally distinct learning mechanisms (i.e., associative vs. propositional learning) or two distinct forms of memory representation (e.g., associative vs. propositional representations). Instead, the observed dissociations may simply reflect differences in the retrieval of stored propositional information, given that (a) responses on implicit and explicit measures differ in terms of their relative speed and (b) fast responses are more likely affected by incomplete retrieval of stored information (e.g., retrieval of *A is related to B* rather than *A prevents B*). Formal modeling techniques for analyzing the multiple causal contributions to overt behavioral responses may help to overcome these ambiguities (e.g., Conrey et al. 2005; Heycke & Gawronski, ms; Hütter & Sweldens 2018).

Finally, the associative-propositional distinction can be unpacked in terms of the *behavioral expression* of evaluative representations. On this view, implicit measures capture the behavioral effects of associatively activated representations, while explicit measures capture the behavioral effects of propositionally validated representations (Gawronski & Bodenhausen, 2006, 2011). Alternatively, it may be that even speeded responses on implicit measures are shaped by quick propositional inferences (De Houwer, 2014). The warrant for distinguishing associative and propositional processes during the expression of mental representations derives from the predictions one can make on its basis. For example, expanding on the predictions derived from dual-process theories, implicit and explicit measures have been used to predict different kinds of behavior (e.g., spontaneous vs. deliberate behavior), the same behavior under different conditions (e.g., under cognitive load), and the behavior of people with varying personality traits (e.g., intuitive vs. deliberative thinking style). (See Brownstein et al. ms, Friese et al., 2008, for extensive discussions of evidence for these predictions.) As we discuss below, these behavior-, context-, and person-related variables should and do moderate the relations between implicit measures and other kinds of behavior (e.g., evaluating CVs).

Traits or States

We have so far reviewed two hypotheses about the nature of the mental constructs underlying performance on implicit measures. Another debated issue is whether performance differences on implicit measures reflect differences in traits or states. Before moving on, however, we note one potential terminological ambiguity in this section. One way of understanding the difference between traits and states has to do with how stable a given construct is over time and across situations. If you have detested sour cream for a long time, and dislike it no matter the dish it's in, then your aversion to sour cream is trait-like. If, however, you like sour cream some days but not others, or when in the company of certain friends but not others, then your feelings toward sour cream are more state-like. Long-standing debates in psychology focus on whether constructs like attitudes, self-esteem, and so on are more trait-like or more state-like (e.g., Fazio, 2007; Schwarz, 2007).

At present, there is significant evidence that the constructs captured by 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 & Payne 2017; Cunningham et al. 2001; Devine et al. 2012; Gawronski et al. 2017). Moreover,

scores on implicit measures appear to be more temporally unstable than individuals' scores on corresponding explicit measures (Gawronski et al. 2017).

Whether this temporal instability is a virtue or a vice of implicit measures depends on one's hypotheses about what they "should" be measuring. If they are intended as measures of spontaneous affective reactions (a type of transient state; Gawronski & Bodenhausen 2006), then we should predict large fluctuation across contexts (one's spontaneous reactions toward images of black people may differ dramatically after watching the film *Black Panther* versus after watching Fox News coverage of "Black Lives Matter"). However, if they are intended to "diagnose" relatively stable traits akin to intelligence or political party affiliation, then we should predict less fluctuation.

One possibility is that at least some of the temporal instability of implicit measures is due to measurement error. In line with this idea, some novel approaches have been successful in improving the temporal stability of implicit measures. 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 (e.g., a group of white people, rather than one white man's face). Embracing the idea that implicit measures capture both traits and states, other researchers have been working to disentangle the relative contributions of trait and state components to performance on implicit measures. For example, Dentale and colleagues (2016) used implicit and explicit measures at multiple occasions and concluded that implicit measures reflect a mix of state (occasion-specific) and trait (temporally stable) components (see also Koch et al. 2014; Lemmer, Gollwitzer, & Banse 2015; Schmukle & Egloff, 2005). Another important consideration has to do with the comparison of measures in different content domains. Implicit measures of political attitudes, for example, appear to be significantly more stable than implicit measures of racial attitudes (Gawronski et al. 2017). This result mirrors findings using explicit measures of political and racial attitudes. Finally, it is worth noting that measuring the transient thoughts and feelings that people have in specific contexts can be 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).

A second way of understanding the difference between traits and states has to do with metaphysics. This approach builds upon long-standing debates in philosophy of mind between dispositionalists and representationalists. The former, which has historical ties to behaviorist and eliminativist approaches to the mind, define concepts like belief, desire, and imagination in terms of dispositions to behave in particular ways. For example, to establish that Kayla believes that it's raining outside, a dispositionalist would point to Kayla's behavior, such as whether she wears a raincoat. (Some dispositionalists would point not only to Kayla's behavior, but to the nexus of her actions, feelings, statements, etc.) Representationalists define concepts like belief in terms of internal, representational structures of the mind. For Kayla to believe that it's raining outside is for Kayla's mind to be in a particular state that corresponds to this belief.

We have described prominent representationalist theories of implicit bias above, some of which suggest that the internal, representational structures are associative and some of which suggest that they are propositional. Other theorists have argued for a dispositional approach to implicit bias, understood in this second, metaphysical sense. For example, Schwitzgebel (2010) argues that implicit measures reveal "in-between beliefs." That is, measures like a race-intelligence IAT demonstrate that people often display part of the dispositional profile associated with racial bias, but not the full dispositional profile. They might unreflectively perceive white people to be more intelligent than black people but fail to report that they believe whites are more intelligent than blacks. Thus, because their dispositional profile is "mixed," people cannot be said to believe, or fail to believe, that white people are more intelligent than black people. Rather, they have an in-between belief.

This broad dispositional approach to attitudes—particularly to the attitude of belief— is hotly contested (see, e.g., Carruthers, 2013). One core challenge for it is to explain what seems to be an important disanalogy between trait-based explanations of action and mental state-based explanations of action. As Carruthers (2013) points out, traits are explanatory as generalizations. To say that Fiona returned the money because she is honest is to appeal to something Fiona would typically do (which perhaps matches a folk-psychological stereotype for honesty). But mental states are explanatory as token causes (of behavior or other mental states). To say that Fiona returned the money because she believes that "honesty is the best policy" is to say that this token belief caused her action in this particular case. Schwitzgebel's broad dispositional approach seems to elide this disanalogy (for replies see Schwitzgebel, 2002, 2013). A related worry has to do with circularity in locating a trait and then using it to predict outcomes. We know a person has a trait (e.g., honesty) because they act in honest ways (e.g., not lying). To then use the trait of honesty to predict whether someone will lie risks confusing description with explanation (Bandura, 1971; Cervone, Caldwell, & Mayer, 2015; Mischel 1968; Payne et al. 2017).

A second trait-based approach for conceiving implicit measures takes them to capture part of the psychological basis of dispositions (Machery 2016, 2017). Machery retains a mental state account of propositional attitudes, according to which, for example, beliefs are stored in and retrieved from a computationally defined, neurally realized "belief box." But according to Machery, attitudes (in the sense that psychologists rather than philosophers use the term, to refer to likings or dislikings) are multitrack dispositions, akin to personality traits. Attitudes do not occur and are not realized in the brain (though they depend in part on brain states). Machery argues that there are various bases that together comprise an attitude. These include feelings, associations, behavior, and propositional attitudes like beliefs. Implicit measures, on this picture, capture one of the many psychological bases of the agent's overall attitude. Explicit questionnaire measures capture another psychological basis of the agent's attitude, behavioral measures yet another basis, and so on. An upshot of this view is that there are no such things as *implicit* attitudes. The reason for this is that the view asserts that attitudes are traits, and traits (like being courageous or extraverted) do not divide into implicit and explicit kinds. A concept like "implicit courageousness" is meaningless, Machery argues. Rather, associations between concepts— which may be implicit in the sense of unconscious or difficult to control— are simply one of the bases of an attitude, as are low-level affective reactions and the like.

The difficulties stated above about whether traits are explanatory apply here too. It is also unclear to what extent both kinds of trait-theories are empirically tractable or whether they are rather efforts at metaphysical housekeeping alone. Moreover, one might question Machery's claim that the implicit/explicit distinction doesn't apply to traits. There are, in fact, implicit measures of personality, and these sometimes dissociate from explicit measures. Moreover, implicit and explicit measures of personality predict different kinds of behavior (e.g., Asendorpf, Banse, & Mücke 2002; Back, Schmukle, & Egloff 2009; Dentale et al. 2016; Grumm & von Collani 2007; Vianello, Robusto, & Anselmi 2010). For example, someone might be courageous in the heat of the moment when they don't have time to think about it, but, when they do have time to think, they second-guess their courageous impulses and act in more cowardly ways.

Finally, Machery's approach is based on his assessment of the apparent psychometric weakness of implicit measures. Specifically, he cites meta-analytic data suggesting that various implicit measures do not correlate with each other, do not predict behavior, and are unstable over time. The trait view, in other words, treats much of the data in research on implicit measures as simply reflecting measurement error. Were implicit measures psychometrically "stronger," we would have reason to think that they capture "full" traits, on Machery's view. But if it is right to consider implicit measures as psychometrically weak, then we have reason to think that what they capture are only the partial basis of traits. To the extent that Machery's account is founded on psychometric concerns rather than higher-level theoretical positions, one challenge is explaining aspects of implicit measures that are not likely attributable to measurement error, such as the high stability of these measures over time when results are aggregated at the county, state, and country-level, and that systematic evaluations have found no evidence of publication bias or a file-drawer problem (see Brownstein et al. ms; Kurdi et al. 2018; Machery 2017; Payne et al. 2017).

Person or Situation

A question closely related to the trait vs. state issue is whether a person's performance on an implicit measure reflects characteristics of the person who is taking the test or characteristics of the situation in which the person is taking the test.[2] Although tensions between person-based and situation-based accounts of attitudes and implicit bias have been recurring for decades (Payne & Gawronski, 2010), situation-based approaches gained new "fuel" by what Payne and colleagues (2017) call the "bias of crowds" model. This model is meant to address five common findings: (1) average group-level scores of implicit bias are very robust and stable; (2) children's average scores of implicit bias are nearly identical to adults' average scores; (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, (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.

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 "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 hearing someone shout "Fire!" will make very different concepts accessible depending on whether

[2] A version of the discussion in this section is found in Brownstein et al. (ms).

one is in a log cabin in the California woods or on a shooting range. Most research on implicit bias 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" (2017, 236), they write. "Instead, most of the systematic variance in implicit biases appears to operate at the level of situations" (2017, 236).

We support Payne and colleagues in calling for a renewed emphasis on situational moderators of the accessibility of the concepts underlying implicit bias. We urge caution, however, in conceiving of implicit measures as measures of situational factors as such. 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., Bandura 1978; Lewin 1936; 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; 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.[3]

In what might be a friendly amendment to Payne and colleagues' bias of crowds model, implicit measures might be conceived of as capturing proxies for 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 (2017) 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 (see Gawronski & Bodenhausen, 2006, 2011). Finally, we note that interactionism of this sort speaks to the question of whether performance on implicit measures points to state-like or trait-like features of the mind (in the first sense of traits and states described in the previous section).

Prediction of Behavior

We do not have space to consider the full range of psychometric issues facing implicit bias research, so we focus here on behavioral prediction.[4] Meta-analytic estimates of average correlations between individuals' scores on implicit measures and measures of behavior have varied, from approximately $r = .14$ to $r = .37$ (Cameron et al. 2012; Greenwald et al. 2009; Kurdi et al. 2018; Oswald et al. 2013). This variety is due to several 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. Nevertheless, according to standard conventions, all of these correlations are considered small to medium. 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).

Expectations about how well a given test performs should be calibrated to the difficulty of measuring what the test aims to capture. How predictively powerful should we expect implicit measures to be? If implicit measures are proxies for unconscious attitudes, for example, then we must consider what expectations we ought to have of the relations between unconscious attitudes and behavior. This is a virtue of resisting the tendency to define implicit measures in terms of unconscious attitudes, associations, etc. Treating conceptualizations of implicit bias as hypotheses about

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

[4] A version of the discussion in this section is found in Brownstein et al. (ms).

what explains test performance allows us to assess those hypotheses in terms of how well they explain the extant data regarding relations between test performance and (other kinds of) behavior.

Since the 1970s, attitude researchers have recognized that any valid attitude measure that ignores person-, context-, and behavior-specific moderators ought to find consistent, positive, but low predictive relations between attitudes and behavior (e.g., Zanna & Fazio 1982). This has been the case with explicit measures, with few exceptions,[5] and it 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. Moreover, when the variables specified by extant theories are considered, relations between test performance and behavior are stronger. For example, Cameron and colleagues (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 = .28). Yet, 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.[6] 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. 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 (2018) find an average correspondence of r = .37 in studies using the most effective IAT designs, for example, those that hew close to Ajzen and Fishbein's (1977) principle of correspondence, which states that attitudes better predict behavior when there is clear correspondence between the attitude object and the behavior in question.

Conclusion

Elsewhere we have made recommendations for improving the science of implicit bias (Brownstein et al. ms; Gawronski, forthcoming). Here we have reviewed extant debates on what implicit measures measure. Toward this end, we advocated for a behavioral approach that treats performance outcomes on implicit measures as behaviors rather than direct indicators of mental constructs. This approach permits a theoretically agnostic assessment of different theories about the nature of and the mental constructs underlying implicit bias. While the data support some theories more than others, there is much more to learn about what implicit bias is, how it can be successfully measured, and what effects it has in the world. We wholeheartedly support continued research on these questions.

Works Cited

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

Anderson, E. (2010). *The imperative of integration*. Princeton, NJ: Princeton University Press.

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

Ayala, S. (2016). Speech affordances: A structural take on how much we can do with our words. *European Journal of Philosophy*, *24*, 879-891.

Ayala-López, S. (2018). A Structural Explanation of Injustice in Conversations: It's about Norms. *Pacific Philosophical Quarterly*, *99*(4), 726–748. https://doi.org/10.1111/papq.12244

Back, M. D., Schmukle, S. C., & Egloff, B. (2009). Predicting actual behavior from the explicit and implicit self-concept of personality. *Journal of Personality and Social Psychology*, *97*, 533-548.

Banaji, M. R. (2011). A vehicle for large-scale education about the human mind. In J. Brockman (Ed.), *How is the internet changing the way you think?* (pp. 392-395). New York: Harper Collins.

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

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

Ikizer, E. G., & Blanton, H. (2016). Media coverage of "wise" interventions can reduce concern for the disadvantaged. *Journal of Experimental Psychology: Applied*, *22*, 135-147.

Blanton, H., & Jaccard, J. (2006). Arbitrary metrics in psychology. *American Psychologist, 61*, 27-41.

Brownstein, M. (2018). *The implicit mind: Cognitive architecture, the self, and ethics*. New York: Oxford University Press.

Brownstein, M. (Forthcoming.) Skepticism about implicit bias. In Beeghly, E. and Madva, A. (Eds.), *An introduction to implicit bias: Knowledge, justice, and the social mind*, New York: Routledge.

[5] See discussion in Brownstein et al. (ms).

[6] 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).

Brownstein, M., & Madva, A. (2012). The normativity of automaticity. *Mind and Language*, 27, 410-434.

Brownstein, M., Madva, A., & Gawronski, B. (Manuscript). Understanding implicit bias: Putting the criticism into perspective.

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.

Carruthers, P. (2013). On knowing your own beliefs: A representationalist account. In N. Nottelmann (Ed.), *New essays on belief: Constitution, content and structure* (pp. 145– 165). Basingstoke: Palgrave MacMillan.

Cervone, D. Caldwell, T. L., & Mayer, N. D. (2015). Personality systems and the coherence of social behavior. In. B. Gawronski & G. V. Bodenhausen (Eds.), *Theory and explanation in social psychology* (pp. 157-179). New York: Guilford.

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

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., & B. K. Payne. (2017). Using groups to measure intergroup prejudice. *Personality and Social Psychology Bulletin*, *43*, 46-59. doi:10.1177/0146167216675331.

Cooley, E., Payne, B. K., Loersch, C., & Lei, R. (2015). Who Owns Implicit Attitudes? Testing a Metacognitive Perspective. *Personality and Social Psychology Bulletin*, *41*(1), 103–115. https://doi.org/10.1177/0146167214559712

Cunningham, W. A., Preacher, K. J., & Banaji, M. R. (2001). Implicit attitude measures: Consistency, stability, and convergent validity. *Psychological Science*, *12*, 163–170. https://doi.org/10.1111/1467-9280.00328

De Houwer, J. (2009). The propositional approach to associative learning as an alternative for association formation models. *Learning & Behavior*, *37*, 1-20.

De Houwer, J. (2014). A propositional model of implicit evaluation. *Social Psychology and Personality Compass*, 8, 342– 353.

De Houwer, J., Gawronski, B., & Barnes-Holmes, D. (2013). A functional-cognitive framework for attitude research. *European Review of Social Psychology, 24*, 252-287.

Dentale, F., Vecchione, M., Ghezzi, V., & Barbaranelli, C. (2016). Applying the latent state-trait analysis to decompose state, trait, and error components of the self-esteem Implicit Association Test. *European Journal of Psychological Assessment.* http://dx.doi.org/10.1027/1015-5759/a000378

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*, 1267-1278.

Dixon, J., & Levine, M. (Eds.). (2012). *Beyond Prejudice: Extending the Social Psychology of Conflict, Inequality and Social Change*. Cambridge University Press.

Dixon, J., Levine, M., Reicher, S., & Durrheim, and 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.

Doris, J. (2002). *Lack of character: Personality and moral behavior*. Cambridge, MA: Cambridge University Press.

Fazio, R. H. (2007). Attitudes as object-evaluation associations of varying strength. *Social Cognition, 25*, 603-637.

Fazio, R. H., Jackson, J. R., Dunton, B. C., & Williams, C. J. (1995). Variability in automatic activation as an unobtrusive measure of racial attitudes: A bona fide pipeline? *Journal of Personality and Social Psychology*, *69*, 1013-1027.

Fleeson, W. (2004). Moving personality beyond the person-situation debate. *Current Directions in Psychological Science, 13*, 83-87.

Forscher, P., Lai, C., Axt, J., Ebersole, C., Herman, M., Devine, P., & Nosek, B. (Manuscript). A Meta-analysis of change in implicit bias.

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*, 285–338. https://doi.org/10.1080/10463280802556958

Gawronski, B. (Forthcoming). Six lessons for a cogent science of implicit bias and its criticism. *Perspectives on Psychological Science.*

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., & Bodenhausen, G. V. (2018). Evaluative conditioning from the perspective of the associative-propositional evaluation model. *Social Psychological Bulletin, 13*, e28024.

Gawronski, B., Brannon, S., & Bodenhausen, G. (2017). The associative- propositional duality in the representation, formation, and expression of attitudes. In R. Deutsch, B. Gawronski, & W. Hofmann (Eds.), *Reflective and impulsive determinants of human behavior* (pp. 103– 118). New York: Psychology Press.

Gawronski, B., Hofmann, W., & Wilbur, C. (2006). Are "implicit" attitudes unconscious? *Consciousness and Cognition, 15,* 485– 499.

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

Gawronski, B., Walther, E., & Blank, H. (2005). Cognitive consistency and the formation of interpersonal attitudes: Cognitive balance affects the encoding of social information. *Journal of Experimental Social Psychology, 41,* 618-626.

Gendler, T. S. (2008a). Alief and belief. *Journal of Philosophy, 105,* 634– 663.

Gendler, T. S. (2008b). Alief in action (and reaction). *Mind and Language, 23,* 552– 585.

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: attitudes, self-esteem, and stereotypes. *Psychological Review, 102*, 4-27.

Greenwald, A. G., & Banaji, M. R. (2017). The implicit revolution: Reconceiving the relation between conscious and unconscious. *American Psychologist, 72*, 861-871.

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. G., Nosek, B. A., Banaji, M. R., & Klauer, K. C. (2005). Validity of the salience asymmetry interpretation of the Implicit Association Test: Comment on Rothermund and Wentura (2004). *Journal of Experimental Psychology: General, 134*(3), 420-425.

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

Grumm, M., & von Collani, G. (2007). Measuring Big-Five personality dimensions with the implicit association test–Implicit personality traits or self-esteem?. *Personality and Individual Differences, 43*, 2205-2217.

Hahn, A., & Gawronski, B. (2018). Facing one's implicit biases: From awareness to acknowledgment. *Journal of Personality and Social Psychology*. http://dx.doi.org/10.1037/pspi00001555.

Hahn, A., Judd, C. M., Hirsh, H. K., & Blair, I. V. (2014). Awareness of implicit attitudes. *Journal of Experimental Psychology: General, 143*, 1369-1392.

Hall, D. L., & Payne, B. K. (2010). Unconscious influences of attitudes and challenges to self-control. In Y. Trope, K. Ochsner, & R. Hassin (Eds.), *Self control in society, mind, and brain* (pp. 221-242). New York: Oxford University Press.

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*, 1–15.

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.

Hermanson, S. (2017). Implicit bias, stereotype threat, and political correctness in philosophy. *Philosophies*, *2*, 12. doi:10.3390/philosophies2020012

Heycke, T. & Gawronski, B. (Manuscript.) Co-occurrence and relational information in evaluative learning: A multinomial modeling approach.

Hofmann, W., Gawronski, B., Gschwendner, T., Le, H., & Schmitt, M. (2005). A meta-analysis on the correlation between the implicit association test and explicit self- report measures. *Personality and Social Psychology Bulletin, 31*, 1369– 1385.

Howell, J. L., Gaither, S. E., & Ratliff, K. A. (2015). Caught in the middle: Defensive responses to IAT feedback among whites, blacks, and biracial black/whites. *Social Psychological and Personality Science, 6*, 373-381.

Howell, J. L., & Ratliff, K. A. (2017). Not your average bigot: The better-than-average effect and defensive responding to Implicit Association Test feedback. *British Journal of Social Psychology, 56*, 125-145.

Hu, X., Gawronski, B., & Balas, R. (2017). Propositional versus dual-process accounts of evaluative conditioning: I. The effects of co-occurrence and relational information on implicit and explicit evaluations. *Personality and Social Psychology Bulletin, 43*, 17-32.

Hütter, M., & Sweldens, S. (2018). Dissociating controllable and uncontrollable effects of affective stimuli on attitudes and consumption. *Journal of Consumer Research, 45*, 320-349.

Johnson, G. (Manuscript). The Structure of Bias.

Jussim, L. 2018. Comment on 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

Kelly, D., & Roedder, E. (2008). Racial cognition and the ethics of implicit bias. *Philosophy Compass*, *3*, 522-540.

Krickel, B. (2018). Are the states underlying implicit biases unconscious?–A Neo-Freudian answer. *Philosophical Psychology, 31*, 1007-1026.

Koch, T., Ortner, T. M., Eid, M., Caspers, J., & Schmitt, M. (2014). Evaluating the construct validity of objective personality tests using a multitrait-multimethod-multioccasion-(MTMM-MO)-approach. *European Journal of Psychological Assessment*, *30*, 208-230.

Kurdi, B., Seitchik, A., Axt, J., Carroll, T., Karapetyan, A., Kaushik, N., Tomezsko, D., Greenwald, A., & Banaji, M. (2018). Relationship between the Implicit Association Test and intergroup behavior: A meta-analysis. *American Psychologist*. http://dx.doi.org/10.1037/amp00000364.

Lee, K. M., Lindquist, K. A., & Payne, B. K. (2017). Constructing bias: Conceptualization breaks the link between implicit bias and fear of Black Americans. *Emotion, 18*, 855-871.

Lemmer, G., Gollwitzer, M., & Banse, R. (2015). On the psychometric properties of the aggressiveness-IAT for children and adolescents. *Aggressive Behavior*, *41*, 84-95.

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

Machery, E. (2016). De-Freuding implicit attitudes. In M. Brownstein & 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 Biases Measure Traits or Situations?. *Psychological Inquiry, 28*(4), 288-291.

Madva, A. (2016a). A plea for anti-anti-individualism: How oversimple psychology misleads social policy. *Ergo*, *3*, 701–728. https://doi.org/10.3998/ergo.12405314.0003.027

Madva, A. (2016b). Virtue, social knowledge, and implicit bias. In M. Brownstein & J. Saul (Eds.), *Implicit bias and philosophy. Volume 1: Metaphysics and epistemology* (pp. 191–215). Oxford: Oxford University Press.

Madva, A. (2017). Biased against bebiasing: On the role of (institutionally sponsored) self-transformation in the struggle against prejudice. *Ergo, 4*, DOI: http://dx.doi.org/10.3998/ergo.12405314.0004.006

Madva, A., & Brownstein, M. (2018). Stereotypes, prejudice, and the taxonomy of the implicit social mind1. *Noûs*, *52*, 611-644.

Mallon, R. (2018). Constructing race: racialization, causal effects, or both? *Philosophical Studies*, *175*(5), 1039–1056. https://doi.org/10.1007/s11098-018-1069-8

Mandelbaum, E. (2016). Attitude, inference, association: On the propositional structure of implicit bias. *Noûs*, *50*, 629-658.

Mann, T. C., & Ferguson, M. J. (2017). Reversing implicit first impressions through reinterpretation after a two-day delay. *Journal of Experimental Social Psychology*, *68*, 122-127.

Mischel, W. (1968). *Personality and Assessment*. Hoboken, NJ: Wiley.

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.

Mitchell, G., & Tetlock, P. E. (2017). Popularity as a poor proxy for utility: The case of implicit prejudice. In Lilienfeld, S. & Waldman, I. (Eds.), *Psychological science under scrutiny: Recent challenges and proposed solutions* (pp 164-195). West Sussex: Wiley.

Moran, T., & Bar- Anan, Y. (2013). The effect of object– valence relations on automatic evaluation. *Cognition and Emotion*, *27*, 743– 752.

Nier, J. A. (2005). How dissociated are implicit and explicit racial attitudes? A bogus pipeline approach. *Group Processes & Intergroup Relations*, *8*, 39-52.

Nosek, B., Smyth, F., Hansen, J., Devos, T., Lindner, N., Ratliff, K., Smith, C., Olson, K., Chugh, D., Greenwald, A., & Banaji, M. (2007). Pervasiveness and correlates of implicit attitudes and stereotypes. *European Review of Social Psychology*, *18*, 36–88.

Olson, M. A., & Fazio, R. H. (2008). Implicit and explicit measures of attitudes: The perspective of the MODE model. In R. E. Petty, R. H. Fazio, & P. Briñol (Eds.), *Attitudes: Insights from the new implicit measures* (pp. 19-63). New York, NY, US: Psychology Press.

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., Burkley, M. A., & Stokes, M. B. (2008). Why do implicit and explicit attitude tests diverge? The role of structural fit. *Journal of Personality and Social Psychology, 94*(1), 16-31.

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,* 277-293.

Payne, B. K., & B. Gawronski. (2010). A history of implicit social cognition: Where is it coming from? Where is it now? Where is it going? In B. Gawronski, and B. Payne (Eds.), *Handbook of implicit social cognition: Measurement, theory, and applications* (pp. 1– 17). New York: Guilford Press.

Payne, B. K., Vuletich, H. A., & Lundberg, K. B. (2017). The bias of crowds: How implicit bias bridges personal and systemic prejudice. *Psychological Inquiry*, *28*, 233-248.

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.

Rentfrow, P. J., Jokela, M., & Lamb, M. E. (2015). Regional personality differences in Great Britain. *PLOS ONE* 10 (3): e0122245. doi:10.1371/journal.pone.0122245.

Rivers, A. M., & Hahn, A. (2018). What cognitive mechanisms do people reflect on when they predict IAT scores?. *Personality and Social Psychology Bulletin*, https://doi.org/10.1177/0146167218799307

Samayoa, J. A. G., & Fazio, R. H. (2017). Who starts the wave? Let's not forget the role of the individual. *Psychological Inquiry*, *28*, 273-277.

Schmukle, S. C., & Egloff, B. (2005). A latent state-trait analysis of implicit and explicit personality measures. *European Journal of Psychological Assessment*, *21*, 100-107.

Schwarz, N. (2007). Attitude construction: Evaluation in context. *Social Cognition, 25,* 638-656.

Schwitzgebel, E. (2002). A phenomenal, dispositional account of belief. *Noûs*, *36,* 249– 275.

Schwitzgebel, E. (2010). Acting contrary to our professed beliefs, or the gulf between occurrent judgment and dispositional belief. *Pacific Philosophical Quarterly*, *91,* 531– 553.

Schwitzgebel, E. (2013). A dispositional approach to attitudes: Thinking outside of the belief box. In N. Nottelmann (Ed.), *New essays on belief: Constitution, content and structure* (pp. 75– 99). Basingstoke: Palgrave MacMillan.

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*

Smith, E. R., & DeCoster, J. (2000). Dual-process models in social and cognitive psychology: Conceptual integration and links to underlying memory systems. *Personality and Social Psychology Review, 4*, 108-131.

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

Van Dessel, P., De Houwer, J., & Smith, C. T. (2018). Relational information moderates approach-avoidance instruction effects on implicit evaluation. *Acta Psychologica, 184*, 137-143.

Van Dessel, P., Gawronski, B., & De Houwer, J. (forthcoming). Does explaining social behavior require multiple memory systems? *Trends in Cognitive Sciences*.

Vianello, M., Robusto, E., & Anselmi, P. (2010). Implicit conscientiousness predicts academic performance. *Personality and Individual Differences, 48,* 452-457.

Wilson, T. D., Lindsey, S. & Schooler, T. Y. (2000). A model of dual attitudes. *Psychological Review, 107,* 101-126.

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/

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.