PSYCHOLOGICAL INQUIRY
2022, VOL. 33, NO. 3, 139–155
https://doi.org/10.1080/1047840X.2022.2106750

Routledge
Taylor & Francis Group

TARGET ARTICLE

Implicit Bias ≠ Bias on Implicit Measures

Bertram Gawronski[a], Alison Ledgerwood[b], and Paul W. Eastwick[b]

[a]University of Texas at Austin, Austin, Texas; [b]University of California, Davis, California

ABSTRACT

People can behave in a biased manner without being aware that their behavior is biased, an idea commonly referred to as implicit bias. Research on implicit bias has been heavily influenced by implicit measures, in that implicit bias is often equated with bias on implicit measures. Drawing on a definition of implicit bias as an unconscious effect of social category cues on behavioral responses, the current article argues that the widespread equation of implicit bias and bias on implicit measures is problematic on conceptual and empirical grounds. A clear separation of the two constructs will: (1) resolve ambiguities arising from the multiple meanings implied by current terminological conventions; (2) stimulate new research by uncovering important questions that have been largely ignored; (3) provide a better foundation for theories of implicit bias through greater conceptual precision; and (4) highlight the broader significance of implicit bias in a manner that is not directly evident from bias on implicit measures.

KEYWORDS

Discrimination; implicit bias; implicit measures; prejudice; stereotyping

If one were to conduct a survey to identify the psychological constructs that nonacademics are most familiar with, *implicit bias* would probably turn out to be a top contender. The construct captures the idea that people may behave in a biased way without being aware that their behavior is biased. Examples of implicit bias can be found under the hashtag #LivingWhileBlack, which includes a long list of mundane, noncriminal activities for which police were called on Black people (e.g., waiting for a friend at Starbucks, moving into an apartment, shopping for prom clothes; see Griggs, 2018). Descriptions of the listed incidents as instances of implicit race bias are based on the assumptions that (1) police would not have been called if the same activities had been performed by a White person and (2) people were unaware that their decision to call the police was influenced by the perceived race of the target. Similar concerns have been raised about implicit gender bias, in that (1) people may show different responses to the actions of a target person depending on the target's gender and (2) people may be unaware that that their responses are influenced by the target's gender.

Although media coverage of such cases has contributed to increased knowledge of the implicit-bias construct among nonacademics and to a surge of organizational trainings designed to bring implicit biases into awareness (see Edwards, 2016; Onyeador, Hudson, & Lewis, 2021; UCnet, 2021), implicit measures of bias deserve enormous credit for providing a tool for the widespread dissemination of the idea that people can be biased without being aware of it (for reviews, see Gawronski & De Houwer, 2014; Lai & Wilson, 2020). By allowing anyone with internet access to complete an Implicit Association Test (IAT; Greenwald, McGhee, & Schwartz, 1998), the Project Implicit website has arguably been more effective in communicating the idea of implicit bias to nonacademic audiences than any article in the popular media or a peer-reviewed academic journal. A central part of this educational endeavor is that many people are rather surprised about the feedback they receive after completing an IAT, often suggesting that they are much more biased than they had assumed prior to completing the task (Howell & Ratliff, 2017; Monteith, Voils, & Ashburn-Nardo, 2001).

A central assumption underlying such educational initiatives is that implicit measures such as the IAT capture implicit bias. Although the meaning and mental underpinnings of responses on implicit measures are the subject of ongoing scientific debates (for a review, see Brownstein, Madva, & Gawronski, 2019), the equation of *implicit bias* and *bias on implicit measures* is rather common and seemingly uncontroversial, which is reflected in descriptions of the IAT and other implicit measures as *implicit bias tests* or *measures of implicit bias* (e.g., Lai & Wilson, 2020; Payne, Niemi, & Doris, 2018). Even the first author of the current article has used these expressions in some of his work (e.g., Gawronski, 2019), essentially equating *implicit bias* with *bias on implicit measures*.

The main goal of the current article is to highlight conceptual and empirical problems with the widespread equation of implicit bias and bias on implicit measures. The central argument is that implicit bias and bias on implicit measures are conceptually and empirically distinct, and that bias on implicit measures should not be treated as an instance of implicit bias. For the sake of precision and brevity, we will use the acronym IB to refer to *implicit bias*,

CONTACT Bertram Gawronski gawronski@utexas.edu University of Texas at Austin, Austin, Texas, USA.

which we define as unconscious effects of social category cues (e.g., cues related to race, gender, etc.) on behavioral responses. IB is contrasted with *explicit bias*, defined as conscious effects of social category cues on behavioral responses, which we refer to with the acronym EB.[1] We will use the acronym BIM to refer to *bias on implicit measures*, which we define as effects of social category cues on behavioral responses captured by measurement instruments conventionally described as implicit. BIM is contrasted with *bias on explicit measures*, defined as effects of social category cues on behavioral responses captured by measurement instruments conventionally described as explicit, which we refer to with the acronym BEM.

Conceptual Ambiguities

What Is IB?

To determine whether IB can be meaningfully equated with BIM, it seems prudent to clearly specify the meaning of the term *implicit* in the two constructs. Although use of the term has been criticized as delusive (Corneille & Hütter, 2020), its meaning is actually very clear when it serves as a qualifier of bias. From a purely behavioral point of view, bias can be defined as the effect of social category cues (e.g., cues related to race, gender, etc.) on behavioral responses (De Houwer, 2019; Payne & Correll, 2020).[2] Expanding on this definition, instances of bias can be described as *explicit* if respondents are aware of the effect of social category cues on their behavioral response. Conversely, instances of bias can be described as *implicit* if respondents are unaware of the effect of social category cues on their behavioral response. Thus, to classify a person's behavioral response toward a target as an instance of IB, one has to demonstrate that (1) the behavioral response is influenced by social category cues and (2) the person is unaware of the effect of the relevant social category cues on their behavioral response. Although determining people's (un)awareness of a given effect can be a methodologically difficult endeavor (see Gawronski & Bodenhausen, 2012; Shanks & St. John, 1994; Sweldens, Corneille, & Yzerbyt, 2014), the definition of IB as a construct is relatively clear and straightforward.

An important aspect of this definition is that it treats IB as a behavioral phenomenon that needs to be explained—not as a "thing" people have that would explain their behavior, as suggested by the way the term is sometimes used in academic and nonacademic writings (for complementary critiques, see Daumeyer, Rucker, & Richeson, 2017; Salter, Adams, & Perez, 2018). De Houwer (2019) discussed several advantages of such a behavioral definition of implicit bias (see also Payne & Correll, 2020). For the purpose of the current analysis, one of the most significant advantages is that it avoids explanatory circularity (De Houwer, Gawronski, & Barnes-Holmes, 2013). Explanations of biased behavior can easily become circular when (1) biased behavior is explained by the proposition that people have IB and (2) IB is inferred from the biased behavior that needs to be explained (for broader discussions of this issue, see Cervone, Shadel, & Jencius, 2001; Fleeson & Jayawickreme, 2021; Gawronski & Bodenhausen, 2015a). A behavioral definition of IB avoids explanatory circularity by clearly distinguishing between IB as a behavioral phenomenon that needs to be explained and the processes and representations claimed to explain the behavioral phenomenon of IB (De Houwer, 2019; De Houwer et al., 2013).[3]

What Is BIM?

In contrast to the unambiguous meaning of the term *implicit* as a qualifier of *bias*, it is much more difficult to determine its meaning in reference to measures (see Corneille & Hütter, 2020; Gawronski & Brannon, 2019), which poses a conceptual challenge to the equation of IB and BIM. To illustrate these difficulties, it is worth starting with a list of measurement instruments that are conventionally described as *implicit*. Table 1 reproduces such a list from a recent Editorial for a Special Issue of the journal *Social Cognition* entitled *Twenty-Five Years of Research Using Implicit Measures* (Gawronski, De Houwer, & Sherman, 2020). What exactly is it that makes these measurement instruments implicit? Based on the meaning of the term *implicit* in IB and the common idea that these instruments can be used to measure IB, one might argue that they capture unconscious effects of social category cues. However, such an interpretation quickly runs into problems if one considers that the IAT, which is arguably the most popular instrument on the list, would not qualify as implicit in this sense, because respondents are typically aware of the effects of social category cues on their responses in the task. Most people quickly notice that their responses are slower and that they make more errors in the bias-incongruent block compared to the bias-congruent block (Monteith et al., 2001; for

[1]Our use of the phrase *effect of social category cues on behavioral responses* should not be taken to mean that the source or cause of a dominant group member's racist or sexist behavior is located within a marginalized group member (see Fields & Fields, 2014). Rather, by defining bias as a behavioral phenomenon, we leave room to ask questions about the explanations for that behavior (as we discuss further in the following pages). We also want to emphasize that the relevant causal force is social category cues at the stimulus level, which may not align with the personal identity of the target (e.g., when someone who identifies as White has stereotypical Afrocentric features).

[2]We use the term *bias* to refer specifically to biases involving social category cues rather than biases in information processing more broadly, the latter of which includes numerous biases that are not directly related to the current question (e.g., hindsight bias, impact bias, etc.).

[3]In line with the distinction between *bias* and *error* (Kruglanski & Ajzen, 1983), the proposed definition treats IB as a behavioral tendency rather than a deviation from a normative criterion of accuracy. A definition referring to accuracy seems problematic, because normative criteria for accurate social perceptions are inherently arbitrary (see Kruglanski, 1989). For example, if implicit race bias is conceptualized with reference to actual similarities and differences between Black people and White people, one would have to specify the relevant populations of Black people and White people, which is inherently arbitrary because there is no a priori basis to determine the relevant population (e.g., Black people living in a particular neighborhood, city, county, state, or country). Similar problems arise for normative conceptualizations in terms of rationality, in that (1) criteria for rational judgments involve a reference to goals and (2) normative propositions about goals are inherently arbitrary. For a more detailed discussion of problems with normative conceptualizations of implicit bias, see De Houwer (2019).

Table 1. Overview of currently available measures that are commonly described as "implicit".

Measurement Instrument	Reference
Action Interference Paradigm	Banse et al. (2010)
Affect Misattribution Procedure	Payne et al. (2005)
Approach-Avoidance Task	Chen & Bargh (1999)
Brief Implicit Association Test	Sriram & Greenwald (2009)
Evaluative Movement Assessment	Brendl et al. (2005)
Evaluative Priming Task	Fazio et al. (1995)
Extrinsic Affective Simon Task	De Houwer (2003)
Go/No-go Association Task	Nosek & Banaji (2001)
Identification Extrinsic Affective Simon Task	De Houwer & De Bruycker (2007)
Implicit Association Procedure	Schnabel et al. (2006)
Implicit Association Test	Greenwald et al. (1998)
Implicit Relational Assessment Procedure	Barnes-Holmes et al. (2010)
Recoding Free Implicit Association Test	Rothermund et al. (2009)
Relational Responding Task	De Houwer et al. (2015)
Semantic Priming (Lexical Decision Task)	Wittenbrink et al. (1997)
Semantic Priming (Semantic Decision Task)	Banaji & Hardin (1996)
Single Attribute Implicit Association Test	Penke et al. (2006)
Single Block Implicit Association Test	Teige-Mocigemba et al. (2008)
Single Category Implicit Association Test	Karpinski & Steinman (2006)
Sorting Paired Features Task	Bar-Anan et al. (2009)
Truth Misattribution Procedure	Cummins & De Houwer (2019)

Table adapted from Gawronski, De Houwer, and Sherman (2020). Reprinted with permission.

related findings regarding the affect misattribution procedure, see Hughes, Cummins, & Hussey, in press; Payne et al., 2013). Hence, if IB is understood as an unconscious effect of social category cues on behavioral responses and the IAT is regarded as one of the most central instruments on the list of implicit measures, it is not feasible to directly apply the notion of IB to the list of measures commonly described as *implicit*.

One alternative might be to interpret the term *implicit* in a procedural manner in the sense of *indirectly measured* (Greenwald & Banaji, 2017; Greenwald & Lai, 2020). According to this interpretation, the defining feature of the instruments listed in Table 1 is that they are all indirect measures, and bias measured via these instruments qualifies as implicit, because it is measured in an indirect rather than direct manner. As explained in detail by Corneille and Hütter (2020), there are numerous problems with such a conceptualization. For the purpose of the current analysis, two problems are especially noteworthy. First, the proposed conceptualization leads to ambiguities in the meaning of IB, in that it can refer to (1) unconscious effects of social category cues on behavioral responses and (2) instances of bias captured by indirect measures. As explained in the previous paragraph with regard to responses on the IAT, the two are not the same, because people can be aware of the effect of social category cues on an indirect measure. Second, the qualifier *indirect* does not provide a clear demarcation between measures commonly referred to as implicit (see Table 1) and other measures that qualify as indirect in the sense that they do not involve self-report, but are not conventionally classified as implicit (for reviews of examples, see Klauer, Voss, & Stahl, 2011; Webb, Campbell, Schwartz, & Sechrist, 1966). Because it is highly unlikely that all currently available indirect measures have the same functional properties, it would seem problematic to expand the use of the term *implicit* to all indirect measures. Moreover, although the notion of IB may be meaningfully related to bias on a subset of all indirect measures, it seems rather implausible that IB would be meaningfully related to bias on all indirect measures. Thus, an interpretation of the term *implicit* as *indirectly measured* is either too broad if it is supposed to refer to all indirect measures or insufficient if it is supposed to refer exclusively to the subset of indirect measures conventionally described as implicit. In the latter case, it would still be unclear which properties would make an indirect measure implicit.

A potential way to address these issues might be to categorize measurement instruments at the mental level rather than the procedural level. Based on the distinction between associative and propositional processes popularized by dual-process theories (e.g., Gawronski & Bodenhausen, 2006; Strack & Deutsch, 2004), the instruments listed in Table 1 are sometimes described as implicit in the sense that their underpinnings are assumed to be associative. Conversely, traditional self-report measures are described as explicit in the sense that their underpinnings are assumed to be propositional. Moreover, whereas the term *associative* is meant to refer to processes and representations involving unqualified mental links between concepts (e.g., mental association between *Aspirin* and *headaches*), the term *propositional* is meant to refer to processes and representations involving the perceived validity of specific relations between concepts (e.g., subjective belief *Aspirin relieves headaches*). There are a number of problems with equating implicit with associative and explicit with propositional (see Corneille & Hütter, 2020).

First, a demarcation of measurement instruments in terms of associative and propositional processes is missing a central aspect of the dual-process theories that inspired this demarcation: the hypothesis that associative and propositional processes operate in an interactive manner rather independently (see Gawronski & Bodenhausen, 2006; Strack & Deutsch, 2004). This hypothesis implies that (1) responses on explicit measures can be influenced by associative processes (see Gawronski & Bodenhausen, 2006, Case 1), and (2) responses on implicit measures can be influenced by

propositional processes (see Gawronski & Bodenhausen, 2006, Case 4). Although the two kinds of influences are assumed to be limited to specific conditions, the possibility that associative and propositional processes can jointly influence responses on either type of measure raises major problems for a definition that equates implicit measures with *associative* and explicit measures with *propositional*.

Second, the amount of evidence for effects of propositional processes on implicit measures has increased to a level that some researchers have rejected the idea of associative processes entirely, advocating for a replacement of dual-process with single-process propositional accounts (e.g., De Houwer, Van Dessel, & Moran, 2020; Kurdi & Dunham, 2020; see also Corneille & Stahl, 2019). An illustrative example is evidence showing that evaluations on implicit measures tend to reflect the specific relation between co-occurring stimuli (e.g., whether A starts or stops an unpleasant stimulus B) rather that their mere co-occurrence (e.g., mere co-occurrence of A with unpleasant stimulus B). If there was a unique mapping between implicit measures and associative processes and between explicit measures and propositional processes, implicit measures should reflect the mere co-occurrence between stimuli, while explicit measures should reflect the specific relation between co-occurring stimuli. The available evidence suggests that such a one-to-one mapping is empirically untenable (for a review, see Kurdi & Dunham, 2020).[4] Whether this evidence supports the proposed rejection of associative processes as a distinct construct is a matter of debate, but it clearly poses a challenge to a definition of implicit and explicit measures in terms of associative and propositional underpinnings.

Another possibility might be to categorize measurement instruments in terms of their functional properties instead of procedural features or underlying mental processes. In line with this idea, De Houwer, Teige-Mocigemba, Spruyt, and Moors (2009) suggested an interpretation of the term *implicit* that is synonymous with the term *automatic*. According to this conceptualization, a measure qualifies as implicit if the to-be-measured psychological attribute (e.g., racial attitude) influences measurement outcomes in an automatic fashion, with the term *automatic* subsuming the independent features of awareness, intentionality, efficiency, and controllability (see Bargh, 1994; Moors, 2016). Because the four features do not overlap, De Houwer et al. (2009) further argued that researchers should specify in which particular sense a measure is claimed to be implicit. Moreover, because any such claims are empirical hypotheses, descriptions of measures as implicit should be supported with appropriate evidence instead of being taken for granted. Applied to the list of instruments conventionally referred to as implicit (see Table 1), two features presumably shared by all instruments are that the measured responses are (1) unintentional and (2) difficult to control. However, similar to a classification of measures as direct versus indirect, numerous other tasks share this characteristic (for a review of examples, see Klauer et al., 2011), and it seems highly unlikely that the broader set of these measures constitutes a sufficiently coherent category that could serve as the basis for a meaningful equation of IB and BIM (see Corneille & Hütter, 2020). Moreover, because unintentional and hard-to-control behavioral effects are not necessarily unconscious (as explained above for the IAT), it seems questionable whether a conceptualization in terms of functional properties can be meaningfully linked to the notion of IB.

Summary

In sum, although the meaning of the term *implicit* is clear and unambiguous with reference to bias, it is rather difficult to determine what exactly makes an implicit measure implicit (see Corneille & Hütter, 2020). Although researchers have tried to address this problem by means of procedural, mental, and functional definitions of the term *implicit* in reference to measures, the proposed definitions either fail to provide a pragmatically useful demarcation of measurement instruments or they involve hypotheses about underlying mental processes that are inconsistent with the available evidence. Thus, despite 25 years of extensive research, the current labeling conventions are still based on conceptually ambiguous lists according to which a measure qualifies as implicit if researchers have described it as implicit in the past (Gawronski, De Houwer, et al., 2020). Because these atheoretical lists fail to identify specific features that make an implicit measure implicit and the lists are not even consistent across publications (see Gawronski, De Houwer, et al., 2020; Greenwald & Lai, 2020; Lai & Wilson, 2020), the conceptual basis for an equation of IB with BIM is rather weak. If it is unclear what constitutes an implicit measure, it also remains unclear what constitutes BIM; and if it is unclear what constitutes BIM, there is no conceptual foundation to equate it with IB. From this perspective, equating IB and BIM seems problematic, because the proposed equation does not have a solid conceptual basis.

Awareness of Bias

Pragmatically oriented researchers may point out that the conceptual ambiguities surrounding the meaning of the term *implicit* with reference to measures did not undermine the enormous progress that has been made by research using implicit measures (see Gawronski, De Houwer, et al., 2020). Moreover, although there seem to be disagreements about the classification of a small number of instruments, there is consensus about the majority of measures described as implicit by extant lists (see Gawronski, De Houwer, et al., 2020; Greenwald & Lai, 2020; Lai & Wilson, 2020). Thus, ignoring the conceptual ambiguities surrounding the meaning of the term *implicit* with reference to measures, a purely empirical argument for the equation of IB and BIM could be made if people are unaware of their BIMs, with BIM

[4] A closely related issue is that, although the majority of implicit measures have been designed with the goal to measure responses arising from associative representations, a subset of implicit measures have been developed to capture responses arising from representations with propositional content (e.g., Barnes-Holmes et al., 2010; Cummins & De Houwer, 2019; De Houwer et al., 2015).

being specified as bias captured by measures for which there is consensus about their description as being implicit. In line with this idea, a common assumption about implicit measures is that they capture unconscious biases that people do not know they have. Although people may become aware of their biases during the completion of an implicit measure (e.g., when people notice that they make more errors in the bias-congruent block compared to the bias-incongruent block of an IAT; see Monteith et al., 2001), people may be unaware of their biases otherwise, in that they are unable to verbally report their biases without the self-insight gained by completing an implicit measure (for a detailed discussion of related ideas, see Hahn & Goedderz, 2020). As noted above, this assumption is central to the use of implicit measures as educational tools to increase people's awareness of their biases.

Empirical evidence for this assumption would render the ambiguities surrounding the meaning of the term *implicit* with reference to measures irrelevant, because it would provide an empirical foundation for the equation of IB and BIM, thereby compensating for the lack of a conceptual foundation. Thus, to the extent that measures conventionally described as implicit capture biases that people are unable to verbally report (at least without the self-insight gained by completing an implicit measure), there would be a solid basis for the equation of IB and BIM. However, a critical question is what would constitute evidence for people's inability to report their biases?

A frequently cited piece of evidence is that correlations between BIM and BEM tend to be rather small overall (for meta-analyses, see Cameron, Brown-Iannuzzi, & Payne, 2012; Hofmann, Gawronski, Gschwendner, Le, & Schmitt, 2005). The central argument is that, if people are unaware of the biases captured by implicit measures, they should be unable to verbally report their biases on explicit measures, which should lead to small correlations between BIM and BEM. Although this deductive inference is logically sound, it does not permit the reverse inference that small correlations between BIM and BEM indicate unawareness of BIM (*fallacy of affirming the consequent*; see Gawronski & Bodenhausen, 2015b). The latter inference would be justified only if there were no other factors that can lead to small correlations between BIM and BEM. Yet, research has identified numerous such factors (see Gawronski, Hofmann, & Wilbur, 2006; Gawronski, LeBel, & Peters, 2007). For the purpose of the current analysis, two factors seem particularly noteworthy. First, many implicit measures suffer from substantial measurement error (Gawronski & De Houwer, 2014; Greenwald & Lai, 2020), which can suppress correlations with other measures, including correlations between BIM and BEM. Indeed, when measurement error is statistically controlled, relations between BIM and BEM are much larger compared to the average relations reported in the literature (Cunningham, Preacher, & Banaji, 2001). Second, rather than being unable to verbally report their biases, people may simply be unwilling to express their biases on an explicit measure (e.g., to maintain a positive view of themselves or their group or to avoid accountability; see Bonam, Nair Das, Coleman, & Salter, 2019; Daumeyer, Onyeador, Brown, & Richeson, 2019; Dunton & Fazio, 1997; Plant & Devine, 1998). Consistent with this idea, relations between BIM and BEM tend to be quite large for individuals low in motivation to control prejudiced reactions and close to zero for individuals high in motivation to control (e.g., Degner & Wentura, 2008; Dunton & Fazio, 1997; Gawronski, Geschke, & Banse, 2003; Payne, Cheng, Govorun, & Stewart, 2005). These findings are difficult to reconcile with the idea that BIM reflects unconscious biases that people are unable to report, but they are consistent with the hypothesis that BIM reflects conscious biases that (some) people are unwilling to report.

Even stronger evidence against the unawareness hypothesis comes from research by Hahn, Judd, Hirsh, and Blair (2014). In a series of studies, participants were asked to predict their BIMs for multiple target groups and then completed several IATs to measure their BIMs for these groups. Counter to the idea that people are unaware of their BIMs, participants showed high accuracy in predicting their IAT scores for the relevant target groups with mean correlations greater than .50 and median correlations around .65. Interestingly, participants were highly accurate in predicting their IAT scores regardless of their prior experience with the IAT, regardless of how much information they received about the IAT in the instructions for the prediction task, and regardless of whether the IAT was introduced as a measure of true beliefs or cultural associations. Moreover, correlations between predicted and actual IAT scores were significantly higher when participants made predictions regarding their own responses than when they made predictions for people in general, which rules out interpretations in terms of naïve theories of bias. Interestingly, participants showed high accuracy in predicting their personal IAT scores although BIM and BEM showed the same small correlations found in prior research (for meta-analyses, see Cameron et al., 2012; Hofmann et al., 2005). Together, these findings provide further support for the idea that people are aware of the biases captured by implicit measures, and that they can be reluctant expressing them on traditional self-report measures (see also Hahn & Gawronski, 2019).

How can these findings be reconciled with anecdotal evidence that many people who take the IAT are rather surprised to learn about their biases (e.g., Banaji, 2011; Krickel, 2018)? If people were aware of their biases, why would they be surprised to learn about them when they take an IAT? To reconcile this apparent contradiction, it is worth noting that surprise reactions can result from a simple mismatch between the naïve metric used by participants to describe the extremity of their biases and the metric used by researchers to convert numeric IAT scores into verbal feedback (e.g., *strong preference for White people compared to Black people*). To the extent that the two metrics do not align, participants may be surprised about their IAT feedback, not because they are unaware of their bias, but because their personal description does not match the description in the feedback they receive. Consistent with this argument, Hahn et al. (2014) found that, although participants were highly accurate in predicting their IAT scores,

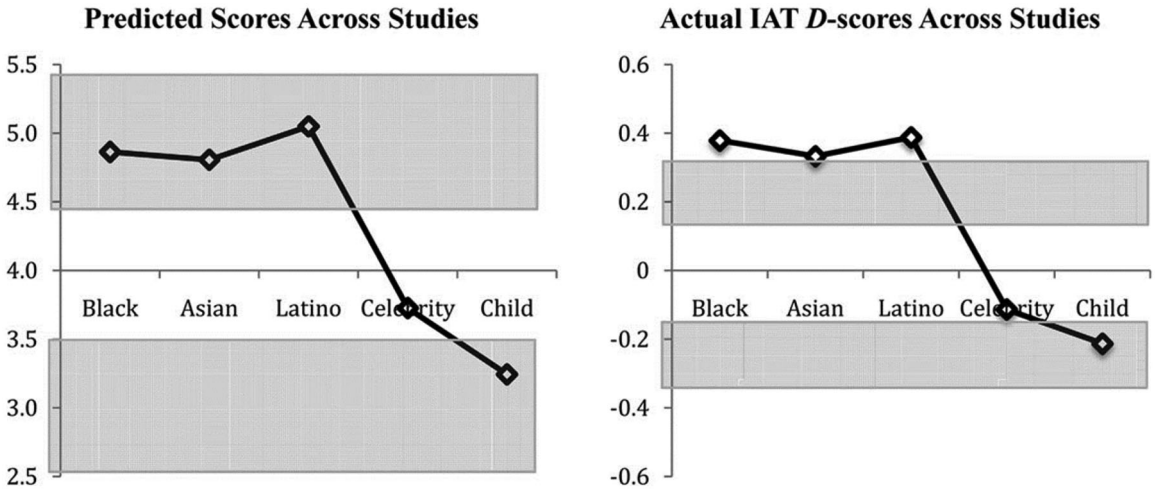

Figure 1. Average IAT score predictions (1–7 scale) and average actual IAT scores. Shaded areas represent the areas in which implicit bias scores would be labeled as "slightly more positive" on the predictions scale or as a "slight preference" according to conventions from the Project Implicit website. Figure adapted from Hahn, Judd, Hirsh, and Blair (2014), reprinted with permission from the American Psychological Association.

their subjective metric to label different levels of bias "stretched" the metric used to convert numeric IAT scores into verbal feedback on the Project Implicit website (see Figure 1). Because labeling conventions for what should be considered a "weak," "moderate," or "strong" bias are entirely arbitrary in the sense that there is no objective basis to treat one metric as "true" and another one as "false" (Kruglanski, 1989), claims that surprise reactions would constitute evidence for the unawareness of BIM are based on a questionable premise and empirically unfounded (see Gawronski, 2019).

In sum, there is no evidence for the idea that people are unable to verbally report their BIM without the self-insight gained by completing an implicit measure. To the contrary, the available evidence suggests that people are well aware of their BIM, and that low correlations between BIM and BEM result from motivational factors influencing the expression of bias on traditional self-report measures. Moreover, surprise reactions in response to IAT feedback seem to be due to misaligned metrics in describing BIM rather than unawareness of BIM. Hence, the lack of a conceptual foundation for the equation of IB and BIM cannot be compensated for with an alternative empirical foundation in the form of evidence that BIM reflect biases that people are unable to verbally report. Because the latter idea is inconsistent with the available evidence, it does not provide a basis for equating IB and BIM.

Relation to Behavior

Although the reviewed issues pose a challenge for a one-to-one equation of IB and BIM, it may be possible to link the two on alternative grounds. One potential basis for such a link would be evidence that BIM is systematically related to IB. In this case, BIM and IB would not be the same constructs, but BIM could still be treated as an indicator of IB even if BIM does not have the same "implicit" properties of IB. That is, people may be aware of BIM and unaware of IB, but BIM may nevertheless predict IB. This idea is related to the notion that, although people may be aware of the thoughts and feelings underlying their responses on implicit measures (see Hahn et al., 2014), they may not be aware of how these thoughts and feelings influence their behavior in other contexts (see Gawronski et al., 2006).

This idea brings up the question about the link between BIM and behavior, which has become the subject of extensive debates. Although the average size of this link differs across meta-analyses (Cameron et al., 2012; Greenwald, Poehlman, Uhlmann, & Banaji, 2009; Kurdi et al., 2019; Oswald, Mitchell, Blanton, Jaccard, & Tetlock, 2013), all of them revealed a small positive relation between BIM and behavior. Yet, at the same time, a recent meta-analysis by Forscher et al. (2019) found no evidence for the hypothesis that changes in BIM would generally lead to corresponding changes in behavior. Together, these findings are consistent with the idea that relations between BIM and behavior may depend on various moderators—and one such moderator could be the difference between IB and EB. That is, BIM may be systematically related to IB but not EB, but behaviors reflecting IB and EB are lumped in meta-analytic reports of average relations between BIM and behavior (including the relations between changes in BIM and changes in behavior in Forscher et al.'s meta-analysis).

In line with this idea, several dual-process theories suggest that BIM should be predictive of spontaneous but not deliberate behavior (e.g., Fazio, 2007; Strack & Deutsch, 2004; see also Dovidio & Gaertner, 2004). Although this hypothesis is consistent with a considerable body of evidence (for a review, see Friese, Hofmann, & Schmitt, 2008), the meaning of the terms *spontaneous* and *deliberate* remains somewhat ambiguous in these studies. In some cases, the distinction seems to refer to behavior performed under conditions of low versus high cognitive elaboration; in some cases, it seems to specify whether the focal behavior

is unintentional or intentional; in some cases, it seems to describe the relative controllability of the focal behavior; and in some cases, it seems to refer to the difference between unconscious and conscious effects on behavior (see Friese et al., 2008). The distinction between unconscious and conscious effects seems closest to the notion of IB, but evidence in this regard is surprisingly scarce, because studies on the relation between BIM and behavior rarely include appropriate awareness checks to confirm the unconscious nature of the effects of social category cues on the focal behavior (see Gawronski & Bodenhausen, 2012). For example, in studies using seating distance as a behavioral criterion (e.g., McConnell & Leibold, 2001), participants may not be aware that their decision to sit further away from a Black person compared to a White person is influenced by the target person's apparent race. Yet, awareness of the impact of racial cues on seating distance is rarely measured in these studies, which is essential for a classification of the focal behavior as an instance of IB.

Based on extant dual-process theories, several meta-analyses on the relation between BIM and behavior have coded the behavioral criterion measures in terms of automaticity features, one being the extent to which the measured behavior is controllable (e.g., Cameron et al., 2012; Greenwald et al., 2009; Kurdi et al., 2019). However, controllability is irrelevant for the notion of IB, because people may be fully aware of the effect of social category cues on their behavior even when the behavior is difficult to control. The only meta-analysis that did code behavioral criterion measures in terms of awareness did not find a systematic difference in the prediction of behavior as a function of awareness (Kurdi et al., 2019). That is, BIM showed the same relation to behavior regardless of whether the focal behaviors qualified as instances of IB versus EB. Thus, although BIM seems to be an indicator of biased behavior in a broader sense, BIM is not uniquely related to IB, which conflicts with the proposed use of BIM as an indicator of IB. To put it differently, if a measure is claimed to be an indicator of IB and this measure is unable to guarantee that the predicted bias is unconscious (and conversely rule out that the predicted bias is conscious), it seems misleading and ill-founded to call this measure an indicator of unconscious bias in the sense of IB.

In sum, the available evidence poses a challenge for the idea that BIM could be treated as an indicator of IB based on a systematic link between the two. Although BIM does seem to be related to IB, this relation is not unique in the sense that BIM shows the same relation to EB. Moreover, if the obtained link between BIM and IB is used to argue for a treatment of BIM as an indicator of IB, the same argument could be made to advocate for a treatment of BIM as an indicator of EB, rendering either of the two treatments arbitrary. Based on these considerations, it seems justified to treat BIM as a potential indicator of biased behavior in a broader sense. However, there is no basis to treat BIM as an indicator of IB, because the link between BIM and biased behavior is not unique to IB. Thus, merely treating BIM as an indicator for IB cannot compensate for the lack of a conceptual and empirical foundation for a direct equation of the two constructs.

The Meaning of *Implicit* in IB

A potential objection against our arguments might be that they are based on a particular interpretation of *implicit* in IB that treats the term as synonymous with *unconscious*. Although this interpretation is widely shared in research on IB/BIM and the field of psychology more broadly (e.g., Greenwald & Banaji, 1995; Roediger, 1990; Schacter, 1987), it seems possible that our conclusions would not generalize to alternative interpretations of the term *implicit* in reference to bias. For example, different from the current interpretation, De Houwer (2019) proposed a behavioral definition of IB in which the term *implicit* subsumes all four features of automaticity instead of exclusively referring to unawareness. According to De Houwer's definition, effects of social category cues on behavioral responses qualify as implicit if they are either (1) unconscious, or (2) unintentional, or (3) efficient, or (4) difficult to control (see Bargh 1994; Moors, 2016). From this perspective, BIM can be regarded as an instance of IB in sense that effects of social category cues captured by implicit measures tend to be unintentional and difficult to control—two functional properties of implicit measures that seem uncontroversial (see above). Moreover, what we have described as IB can also be regarded as an instance of IB in the sense that effects of social category cues on behavioral responses are unconscious. Thus, one might argue that what appears to be an unsurmountable problem in our preceding analysis is actually limited to a narrow interpretation of *implicit* as unconscious, in that the identified problem disappears if one adopts a broader interpretation of *implicit* as automatic.

We argue that such a broader interpretation merely conceals the identified problem without actually resolving it. A major downside of treating the term *implicit* as synonymous with *automatic* is that it creates conceptual ambiguity. In line with this concern, De Houwer (2019) suggested that researchers should specify in which particular sense an observed instance of bias is claimed to be automatic: is it unconscious, unintentional, efficient, or difficult to control? Using the term *implicit* for all of these features without further specification does not mean that all instances of bias classified as implicit have the same properties. After all, effects of social category cues that are unintentional and/or difficult to control are not necessarily unconscious. Because different features of automaticity do not overlap, each hypothesized feature requires independent empirical confirmation (see Bargh, 1994; Moors, 2016). Thus, the uncontroversial idea that BIM is unintentional and difficult to control does not imply unawareness, and the available evidence clearly speaks against the latter hypothesis (see above). From this perspective, a broader interpretation of *implicit* that is synonymous with *automatic* merely conceals the problem via undifferentiated use of the same term for conceptually distinct features.

Table 2. Potential interpretations of the term *implicit* with reference to *bias on implicit measures* and problems associated with a given interpretation for the common equation of *implicit bias* and *bias on implicit measures*.

Interpretation	Problem
Bias on implicit measures reflects unconscious (in contrast to conscious) effects of social category cues on behavioral responses.	People tend to be aware of effects of social category cues on their responses on prominent implicit measures.
Bias on implicit measures reflects indirectly measured (in contrast to directly measured) effects of social category cues.	List of available indirect measures is too broad to form a sufficiently coherent category, and not all indirect measures are deemed implicit.
Bias on implicit measures is an outcome of associative (in contrast to propositional) processes.	Responses on implicit measures have been shown to be influenced by propositional processes.
Bias on implicit measures is an outcome of automatic (in contrast to controlled) effects of social category cues on behavioral responses.	Although effects captured by many implicit measures are unintentional and difficult to control, unintentionality and uncontrollability do not overlap with unawareness.
Bias on implicit measures reflects biases that people are unaware of (in contrast to biases that people are aware of).	People can predict their biases on implicit measures with a high degree of accuracy.
Bias on implicit measures is a unique predictor of unconscious (in contrast to conscious) effects of social category cues on behavior.	Predictive relation between implicit measures and behavior is not unique to unconscious effects of social category cues, but also includes conscious effects.

Of course, one could try to avoid conceptual ambiguity by specifying the particular sense in which an observed instance of bias is claimed to be automatic (see De Houwer, 2019). However, such a strategy does not resolve the problem either, because it makes the issue reappear under different terminology (e.g., unintentional bias $\neq$ unconscious bias). Because people can be fully aware of unintentional, hard-to-control effects of social category cues (as is the case for BIM), such instances of bias do not capture the idea that people can behave in a biased way without being aware that their behavior is biased, leaving the latter phenomenon in the same "blind spot" where it has been due to the equation of IB and BIM. Moreover, because unconscious biases have the potential to cause social harm in ways that are fundamentally different from conscious biases that are unintentional and hard-to-control, the two kinds of biases arguably require different strategies to combat their harmful effects. Thus, redefining *implicit* to mean *automatic* or *unintentional and hard-to-control* (rather than *unconscious*) not only leaves the identified problems unresolved; it can also lead to ill-founded conclusions about how to tackle the harmful effects of unconscious biases.

Moving Forward

The preceding analysis suggests that, if IB is defined as an unconscious effect of social category cues on behavioral responses, there is no basis to equate IB with BIM (see Table 2). Based on this conclusion, it seems prudent to clearly separate IB from BIM and not presume that empirical evidence about the properties of BIM provides information about the properties of IB. The latter question requires direct investigations of IB, and this endeavor cannot be accomplished by using BIM as a proxy for IB.

Identifying IB

To investigate IB, one first has to establish the presence of bias (i.e., is there a causal effect of social category cues on behavior?) and then determine whether the observed bias is implicit or explicit (i.e., are people aware of the causal effect of social category cues on their behavior?). Although it can be difficult to unambiguously establish causal effects of social category cues in natural settings (because social category cues are often confounded with multiple other factors), carefully controlled lab experiments are a valuable tool to determine the extent to which behavioral responses are influenced by social category cues. Yet, even in carefully controlled lab experiments, determining the unconscious nature of such causal effects can be a challenging task (see Gawronski & Bodenhausen, 2012; Shanks & St. John, 1994; Sweldens et al., 2014). To the extent that participants are able to accurately report an actually existing effect of social category cues on their responses, there would be clear evidence for awareness of bias (e.g., Hahn et al., 2014). However, a mismatch between self-reported and actual effects of social category cues is insufficient for inferring unawareness, because self-reports of causal effects of social category cues could be distorted by self-presentational concerns even when people are fully aware of their biased behavior (Dunton & Fazio, 1997; Plant & Devine, 1998; see also Bonam et al., 2019; Daumeyer et al., 2019).

Drawing on extant theories of bias correction (Strack & Hannover, 1996; Wegener & Petty, 1997; Wilson & Brekke, 1994), a potential strategy to address this ambiguity is to eliminate lack of motivation and lack of ability for bias correction as potential explanations of uncontrolled effects of social category cues. To the extent that people are highly motivated to control effects social category cues on their behavior and also the have ability to do so, a plausible explanation of persistent effects of social category cues would be lack of awareness, in that people would control effects of social category cues if they were aware of these effects (e.g., Gawronski et al., 2003).

If empirical evidence confirms the implicit nature of bias, important follow-up questions pertain to the causes of IB, its consequences, and its underlying processes. Regarding the last question, Gawronski, Ledgerwood, and Eastwick (2020) discussed the roles of two potential mechanisms that have received considerable empirical attention: (1) biased interpretation and (2) biased weighting (see also Bodenhausen, 1988).

Mechanisms Underlying IB

Biased interpretation occurs when people construe the same information about a target differently depending on social

category cues associated with the target. Illustrative examples are the mundane, noncriminal activities under the hashtag #LivingWhileBlack for which police were called on Black people (see Griggs, 2018). The broader concern in these incidents is that the individuals who called the police construed the mundane activities of Black people as suspicious and threatening, and that they would not have construed the activities in the same way if the targets had been White people. Such effects of social category cues on the interpretation of target information have been demonstrated in numerous experimental studies (e.g., Darley & Gross, 1983; Duncan, 1976; Gawronski et al., 2003; Hugenberg & Bodenhausen, 2003; Kunda & Sherman-Williams, 1993; Sagar & Schofield, 1980). For example, in a study by Hugenberg and Bodenhausen (2003), participants perceived the same neutral facial expressions as more hostile when the target was Black than when the target was White (see also Bijlstra, Holland, Dotsch, Hugenberg, & Wigboldus, 2014; Hutchings & Haddock, 2008). Moreover, consistent with the hypothesis that such biased interpretations occur outside of awareness (e.g., Fazio, 1990; Trope, 1986), some studies obtained effects of social category cues on the interpretation of target information even when participants were motivated and able to respond in an unbiased manner (e.g., Gawronski et al., 2003).

Biased weighting occurs when people weigh the same information about a target differently depending on social category cues associated with the target. An illustrative example is the biased weighting of credentials in hiring decisions. The broader concern in such cases is that decision-makers may sometimes weigh credentials in a manner that merely serves to rationalize a preexisting preference instead of generating a preference based on the candidates' credentials. For example, in a study by Norton, Vandello, and Darley (2004), participants who were asked to review job application materials of a man and a woman showed a general preference for the man regardless of the candidates' credentials. Specifically, participants preferred the man when (1) the man had more work experience but less education than the woman did, but also when (2) the man had more education but less work experience than the woman did. In both cases, participants justified their decisions with whatever qualification made the man superior to the woman, suggesting that they weighed the candidates' credentials in a manner that served to rationalize a preexisting preference for the man (see also Hodson, Dovidio, & Gaertner, 2002; Uhlmann & Cohen, 2005). Further research found that participants' self-perceptions of objectivity in their hiring decision were associated with greater (rather than smaller) bias (Uhlmann & Cohen, 2005). This relation is consistent with the hypothesis that biasing effects of differential weighting can occur outside of awareness, but the observed relation could also be driven by self-presentational concerns.

Although biased interpretation and biased weighting are both well-established mechanisms leading to biased behavior, it is worth noting that evidence for their unconscious operation is still relatively scarce. Because both mechanisms may involve either conscious or unconscious effects of social category cues, future research would be helpful to establish the conditions under which biased interpretation and biased weighting influence judgments and decisions outside of awareness. In instances involving unconscious effects of social category cues, the respective findings would provide valuable information on the mental processes that can lead to IB. In such cases, unconscious effects of social category cues on behavioral responses would be the behavioral phenomenon that needs to be explained and biased interpretation and biased weighting would be two potential mechanisms that explain the behavioral phenomenon of IB (see De Houwer et al., 2013).[5]

Expanding on evidence for the roles of biased interpretation and biased weighting, an interesting follow-up question is what predicts the outcome of the two mechanisms and their conscious versus unconscious operation. Although IB should not be equated with BIM, studies on this question might suggest a systematic link with BIM. For example, research may suggest that some instances of IB result from biased interpretation of target information, and that biased interpretations in these cases are systematically related to BIM but not BEM (e.g., Gawronski et al., 2003).[6] Similarly, research may suggest that biased weighting can be involved in both EB and IB, in that biased weighting may occur either consciously or unconsciously. Moreover, whereas conscious operation of biased weighting may be systematically related to BEM but not BIM, unconscious operation of biased weighting may be systematically related to BIM but not BEM. Although these questions naturally follow from a clear separation of IB and BIM, empirical evidence for these links is surprisingly scare, presumably due to the common conflation of IB and BIM.

Bias Intervention

Another important question for future research concerns the factors that influence IB. In addition to providing valuable information about the sources of IB, research on this question has important implications for interventions to reduce IB, including implicit-bias trainings in organizational contexts (Carter, Onyeador, & Lewis, 2020; Onyeador et al., 2021). Drawing on the common equation of IB and BIM, numerous studies have tested the effectiveness of various interventions in reducing BIM, assuming that any such reductions are associated with corresponding reductions in discriminatory behavior (e.g., Lai et al., 2014). However, the latter assumption conflicts with evidence suggesting that intervention-related changes in BIM are not generally associated with corresponding changes in discriminatory behavior (Forscher et al., 2019) and, conversely, intervention-related changes in discriminatory behavior are not generally

[5]It is worth noting that our discussion of potential mechanisms underlying IB is not meant to be exhaustive, in that IB could result from multiple other mechanisms besides biased interpretation and biased weighting.

[6]Consistent with this hypothesis, Hugenberg and Bodenhausen (2003) found that biased interpretations of neutral facial expressions were associated with BIM but not BEM. However, their study did not include any data that could confirm unawareness, rendering interpretations in terms of IB premature.

associated with corresponding changes in BIM (Forscher, Mitamura, Dix, Cox, & Devine, 2017). Although these studies have focused on discriminatory behavior more broadly rather than IB in particular, they are consistent with the current argument that IB should not be equated with BIM. The broader implication of this argument for bias intervention is that research on BIM should not be used to draw inferences about how we can reduce IB. Instead, interventions designed to reduce IB need to be evaluated for their effectiveness in reducing actual instances of IB. Evidence for intervention-related reductions in BIM is not suitable to address this question, because IB and BIM are not the same.

One important question in this context is how one could help people identify effects of social category cues in order to provide a basis for bias correction (see Strack & Hannover, 1996; Wegener & Petty, 1997; Wilson & Brekke, 1994). Research on bias correction suggests that naïve realism might be a major obstacle in this endeavor. Naïve realism refers to the phenomenon that people treat their subjective perceptions as direct reflections of objective reality (Ross & Ward, 1996), which can undermine efforts to correct for IB. Moreover, research on motivated ideologies suggests that people may not be willing to acknowledge their biases for multiple reasons (Neville, Awad, Brooks, Flores, & Bluemel, 2013; Wellman, Wilkins, Newell, & Stewart, 2019), which can similarly undermine efforts to correct for IB. For example, in cases involving biased interpretations of mundane activities by Black people, a person may state that they should call the police on anyone who is trying to break into a house regardless of whether the target is White or Black. However, they may not realize that they are interpreting a target's behavior as "trying to break into a house" only when the target is Black, and that they would not interpret the same behavior in this way if the target was White. Similarly, in cases involving biased weighting of work-related credentials by a man and a woman, a person may rationalize their preference for hiring the man based on unique credentials of that candidate. However, they may not realize that their justification is arbitrary in the sense that they would express the same preference with a different justification if the credentials of the two candidates were reversed. Thus, in both cases, people may deny that social category cues had any influence on their responses and refer primarily to their perceptions of the target, without realizing that their perceptions of the target are influenced by social category cues (see Dovidio & Gaertner, 2004; Ledgerwood, Eastwick, & Gawronski, 2020). Implicit-bias trainings that educate people about IB as a phenomenon may be an important first step, but without practical advice on how to identify IB in specific situations and without interventions that tackle the motivated processes that counteract acknowledgement of bias, such trainings will likely be ineffective in combatting IB (see also Carter et al., 2020). Moreover, even when such broader interventions are available, interventions at the individual-level will most likely have to be supplemented with changes at the structural level of decision environments to tackle IB more effectively (Onyeador et al., 2021). Although some recommendations on these issues can be derived from the literature on bias correction (see Gawronski, Ledgerwood, et al., 2020) and racial identity development (see Helms, 1997), we still know surprisingly little about the most effective ways to reduce IB. Research investigating the effectiveness of bias interventions in reducing BIM are not suitable to address these questions, because BIM is not the same as IB.

Psychometrics of IB

Solid answers to these questions require valid and reliable paradigms to study IB. Yet, to some readers, our rejection of BIM as a paradigm to study IB may sound like advocacy for going back to the modal practice prior to Greenwald and Banaji's (1995) seminal review of implicit social cognition research. A major problem identified by Greenwald and Banaji was that many studies in this area relied on poorly validated ad-hoc measures with questionable psychometric properties. This situation changed significantly with the development and widespread use of implicit measures. Although implicit measures differ considerably in terms of their psychometric properties (see Gawronski & De Houwer, 2014; Greenwald & Lai, 2020), research using implicit measures has paid much more attention to these issues compared to implicit social cognition research conducted prior to the publication of Greenwald and Banaji's (1995) article. Does our rejection of BIM as a paradigm to study IB move us back to the time of poorly validated ad-hoc measures with questionable psychometric properties?

To be clear, we fully agree with Greenwald and Banaji's (1995) conclusion that solid research on IB requires properly validated instruments with satisfactory psychometric properties. Such instruments still do not exist more than a quarter century later. Moreover, although the development of implicit measures has inspired an enormous amount of research on BIM, this research provides no information about IB if IB is understood as an unconscious effect of social category cues on behavioral responses. Based on these observations, we would argue that research on IB has made very little (if any) progress since Greenwald and Banaji (1995), primarily due to the shift in research foci from IB to BIM that resulted from the equation of the two constructs. Yet, in contrast to the stagnation of research on IB, there has been considerable progress in research on unconscious processes more broadly (see Gawronski & Bodenhausen, 2012; Hahn, & Goedderz, 2020; Newell & Shanks, 2014; Sweldens et al., 2014). This research has provided valuable insights into how one should (and should not) study unconscious processes, and these insights should be quite helpful if we accept the challenge of going back and starting over where the predominant focus on BIM led to a stagnation of research on IB. As we discuss in the final section, IB likely plays a significant role in the perpetuation of disparities at the societal level, a topic many researchers in this area care about. However, our understanding of these issues is still very limited due to the scarcity of research on IB (which stands in contrast to the massive amount of research on BIM).

Links to Societal Disparities

For many psychologists, the interest in BIM and IB is rooted in the idea that empirical findings regarding these constructs may help them understand and tackle disparities at the societal level. Thus, an important question arising from the distinction between BIM and IB is how societal disparities may be related to BIM and IB, respectively.

Societal Disparities and BIM

Regarding a potential contribution of BIM to societal disparities, it seems unlikely that BIM functions as a direct cause of disparities at the societal level, but it may be indirectly related to societal disparities via discriminatory behavior that is associated with BIM. Although average correlations between BIM and discriminatory behavior vary across meta-analyses (Cameron et al., 2012; Greenwald et al., 2009; Kurdi et al., 2019; Oswald et al., 2013), the average correlations obtained in these meta-analyses tend to be relatively small overall. Some have argued that statistically small associations between BIM and discriminatory behavior at the individual level can still have large effects at the societal level if they affect many people simultaneously or repeatedly affect single persons (Greenwald, Banaji, & Nosek, 2015), but in the absence of direct evidence for the background assumptions underlying such claims, the presumed role of BIM as a cause of societal disparities is still unclear (see Mallon, 2021).

Regarding a potential contribution of societal disparities to BIM, some researchers suggested that BIM reflects biases in the environment of the person completing an implicit measure rather than individual biases of the person. For example, in their bias-of-crowds model, Payne, Vuletich, and Lundberg (2017) argued that (1) societal disparities influence the situational accessibility of bias-related concepts and (2) BIM is shaped primarily by situationally rather than chronically accessible concepts. These hypotheses reconcile three sets of paradoxical findings in the literature on BIM. First, they explain how BIM can be widespread and robust on average (Nosek et al., 2007), yet highly unstable over just a few weeks at the individual level (Gawronski, Morrison, Phills, & Galdi, 2017). Second, they explain how BIM can be highly stable across age starting from early childhood (Dunham, Baron, & Banaji, 2008) despite being highly unstable over just a few weeks (Gawronski et al., 2017). Third, they explain why aggregate scores of BIM at the regional level show strong associations with societal disparities (Hehman, Calanchini, Flake, & Leitner, 2019), although meta-analytic associations between BIM and discriminatory behavior at the individual level tend to be relatively small overall (Cameron et al., 2012; Greenwald et al., 2009; Kurdi et al., 2019; Oswald et al., 2013). According to the bias-of-crowds model, robust average levels of bias over time and across age groups reflect the relative stability of disparities at the societal level, while short-term fluctuations at the individual level reflect variations in concept accessibility driven by incidental features of a person's context. Moreover, strong associations between societal disparities and aggregate scores of BIM at the regional level reflect a causal effect of situational factors on the accessibility of bias-related concepts, while the reverse effect of accessible concepts on discriminatory behavior at the individual level may be regarded as spurious, at least from the predominantly situational perspective of the bias-of-crowds model.

Although the bias-of-crowds model deserves credit for taking a step toward connecting work on systemic and institutional racism (e.g., Jones & Carter, 1996; Salter et al., 2018) with research on BIM, the model has also been the target of criticism. For the purpose of the current analysis, two critical arguments seem especially relevant. First, the model is unable to explain why the temporal stability of BIM at the individual level is substantially higher when the situational context is controlled in a way that is meaningfully related to the focal target group (Gschwendner, Hofmann, & Schmitt, 2008). According to the bias-of-crowds model, any individual differences in BIM obtained under such conditions should be measurement noise, which should decrease (rather than increase) the temporal stability of BIM at the individual level (Gawronski & Bodenhausen, 2017). Second, the set of paradoxical findings can also be explained as the product of noisy measurement of BIM and reduction of measurement noise via aggregation (Connor & Evers, 2020). Whereas aggregation of data across individuals isolates situation-related variance by eliminating effects of person-related factors (e.g., Payne et al., 2017), aggregation of data across situations isolates person-related variance by eliminating effects of situation-related factors (e.g., Ajzen, 1987). Either of the two approaches can be helpful to eliminate variance caused by factors that are unrelated to a focal research question. However, both approaches are limited in their ability in capturing the complex role of person-by-situation interactions, as reflected in the finding that individual differences in BIM are much more stable over time when the situational context is systematically controlled (Gschwendner et al., 2008). Thus, although it is possible that societal disparities contribute to BIM, the idea that BIM could be interpreted as a direct indicator of societal disparities seems problematic on both conceptual and empirical grounds.

Societal Disparities and IB

Regarding the relation between IB and societal disparities, it seems likely that unconscious effects of social category cues contribute to disparities at the societal level. If consequential decisions regarding hiring, promotions, housing, policing, criminal sentencing, etc. are influenced by social category cues outside of decision-makers' awareness, and such influences occur at a sufficiently high rate, they will surely lead to systematic disparities at the societal level. However, empirical evidence on the link between individual-level psychological processes and macro-level societal outcomes is still scarce, presumably due to the need for interdisciplinary approaches to tackle this important issue (e.g., Hailey, 2020). Future research combining methods from multiple

disciplines within the social sciences (e.g., psychology, sociology, economics) would help to fill this gap.

Regarding the reverse link, empirical findings in the broader literature lend support to the idea that societal disparities may lead to IB. A central aspect of societal disparities is that they involve a history of group-related differences in social roles and status positions (e.g., organizational policies and national propaganda campaigns in the U.S. pushed women out of the traditionally male workplace following World War II; Honey, 1984). A considerable body of evidence suggests that people readily infer corresponding dispositions from role-constrained behaviors without taking the impact of situational role-constraints into account (Gilbert & Jones, 1986; Humphrey, 1985; Ross, Amabile, & Steinmetz, 1977), a phenomenon known as the *fundamental attribution error* or *correspondence bias* (for a review, see Gawronski, 2004). Similar attributional biases have been found at the group level (Allison & Kerr, 1994; Allison & Messick, 1985; Mackie & Allison, 1987; Worth, Allison, & Messick, 1987), suggesting that group-related asymmetries in social roles and status positions can shape mental representations of social groups (Diekman & Eagly, 2000). To the extent that these representations influence the interpretation and weighting of information about group members in a conscious manner, societal disparities may contribute to EB. If such effects occur unconsciously, societal disparities could be a source of IB. To the extent that IB conversely contributes to societal disparities, these mechanisms suggest a loop of mutually enhancing factors, in which historically rooted disparities contribute to IB and then IB perpetuates these disparities. Although findings in the broader literature are consistent with such a conclusion, systematic investigations regarding the mutually reinforcing nature of IB and societal disparities are still lacking. Future research on these questions would be helpful to fill this gap.

Societal Disparities Without Bias

Although it seems likely that both IB and EB contribute to disparities at the societal level, it is worth noting that their shared definition of *bias* as the causal effect of social category cues on behavioral responses does not capture another important aspect of societal disparities known as *disparate impact*. U.S. labor law specifies disparate impact in terms of practices that adversely affect a particular social group even when the practices themselves are formally neutral (see Title VII of the Civil Rights Act of 1964). Although such practices can lead to systematic discrimination and perpetuate historically rooted disparities, they are not captured by the psychological concept of *bias* if it is defined as the causal effect of social category cues on behavioral responses.

To illustrate this issue, consider the fact that, for a considerable period in the history of the United States, institutionalized housing discrimination locked African Americans out of building wealth through home ownership (Rothstein, 2017). Now imagine a current-day mortgage lender who uses formally neutral criteria indicating wealth (e.g., bank balance) to determine a buyer's pre-approved loan amount. In terms of a definition of bias as the effect of social category cues, the lender would not be engaging in biased behavior because race is not a causal factor influencing their decision. Nevertheless, such decision criteria have disparate impact in the sense that they systematically disadvantage African Americans due to the history of racial discrimination in housing. Research focusing exclusively on bias, as defined here, is unable to capture such systemic sources of societal disparities.

This example illustrates the chasm between the bias concept in psychological research and the real-world disparities that many psychologists endeavor to solve. Thus, understanding and reducing disparities at the societal level requires a broader, historically-rooted approach that goes beyond a purely psychological conceptualization of bias (Hooks, 2003; Salter et al., 2018). Focusing specifically on the goals of implicit-bias trainings, it may not be enough to raise awareness of the effects of social category cues and to train decision-makers to control effects of social category cues. Rather, interventions would need to include changes at the system level to shift decision-makers' focus away from criteria that have historically been easier to acquire for some social groups than others (Ledgerwood et al., 2021; Salter et al., 2018).

Situating psychological treatments of bias in the context of historical and systemic disparities raises a number of complex issues, especially when we consider the many ways that real-world decisions can be influenced by social category cues. To illustrate these complexities, consider the following example: when selecting a person to fill a position, Avery selects a strong White candidate with Credential X over a similarly strong Black candidate without Credential X. The distinction between EB, IB, and disparate impact suggest three potential scenarios involving Avery's preference for the White candidate. First, Avery might want to hire a White person and is consciously using Credential X to justify this preference in a seemingly race-neutral manner. In this case, Avery can be said to exhibit EB in the sense that (1) social category cues influenced Avery's decision and (2) Avery is aware of the influence of social category cues. Second, Avery might have a preference for the White candidate with Credential X, but may not realize that this preference is rooted in the candidate's race rather than Credential X. In this case, Avery can be said to exhibit IB in the sense that (1) social category cues influenced Avery's decision and (2) Avery is unaware of the influence of social category cues. Finally, Avery might have a preference for the White candidate with Credential X and would have chosen whichever candidate had Credential X regardless of its correlation with race. In this case, Avery's decision would not be biased according to the definition used here, but the decision would still have disparate impact if Credential X is generally more common for White than Black candidates (e.g., due to historical discrimination).

One could imagine constructing a tightly controlled lab experiment to tease apart these three cases. By orthogonally manipulating social category cues and their correlation with

a particular attribute (e.g., having Credential X), it is possible to determine whether the attribute has the same effect regardless of its correlation with social category cues in the experiment or whether the effect of the attribute on a decision changes as a function of its correlation with social category cues (e.g., Norton et al., 2004). To the extent that there is evidence for bias (i.e., a causal effect of social category cues that is independent of the attribute), awareness of bias could be probed by ensuring high motivation and high ability to correct for bias. Although carefully designed manipulations of this kind permit identification of implicit bias in the lab, tight control of these factors is virtually impossible for behavior and decisions in real-world contexts, where correlates of social category cues are often deeply historically entrenched, and people learn about (and use and justify) these correlates gradually over time (see Mueller, 2017; Salter et al., 2018).

Given these ambiguities, it would be fallacious to infer that the use of correlated attributes in real-world decisions is generally unbiased. After all, reliance on a seemingly neutral decision criterion may be rooted in its correlation with social category cues. Moreover, even if a correlated attribute is used independent of its correlation with social category cues, it can still perpetuate historical disparities in a manner that evades a purely psychological conceptualization of bias. Thus, although EB and IB are clearly important for understanding and tackling societal disparities, they are insufficient because a psychological conceptualization bias as the effect of social category cues does not capture systemic factors involving disparate impact.

Conclusions

Some critics have dismissed the idea of IB based on extant controversies surrounding implicit measures, and such skeptical conclusions can be found in both academic (e.g., Schimmack, 2021) and nonacademic (e.g., MacDonald, 2017) writings. Such criticism does not seem surprising given the common equation of IB and BIM. However, it is premature in light of the current arguments for why IB should not be equated with BIM. Even if all critiques of BIM research were valid, a complete dismissal of this research would have no direct implications for the construct of IB, which is conceptually and empirically distinct from BIM.

Although it may be difficult to break the terminological habit of labeling the outcomes of implicit measures as IB, a strict distinction between IB and BIM has several advantages. First, by offering a clear and unambiguous definition of IB, it resolves conceptual ambiguities arising from the multiple meanings implied by current terminological conventions. Second, by highlighting several important questions that have been largely ignored, it has the potential to stimulate new empirical research. Third, by promoting conceptual precision in the interpretation of data, it provides a better foundation for theories of IB. Fourth, by conceptualizing IB in a manner that permits direct applications to meaningful behavior in natural settings, it highlights the broader significance of IB in a manner that is not directly evident from BIM.

In response to the identified problems arising from the equation of IB and BIM, proponents of the equation may suggest that it might be easier to adopt a terminological practice that refers to BIM as IB by mere convention, using a theoretically agnostic list of instruments that does not invoke any conceptual or empirical assumptions about the instruments on that list. Although such a terminological convention would address the identified problems of equating IB and BIM, it would downgrade IB research to a level where it becomes equivalent to investigations of responses on computerized tasks. Without strong assumptions regarding a systematic link between such responses and meaningful behavior, such an approach would severely reduce the value of IB research for understanding biased behavior and social discrimination.

Of course, critics of implicit measures may draw the same conclusion for BIM research in general, even when the distinction between IB and BIM is taken seriously. Acknowledging the possibility of such a conclusion, the current proposal can be interpreted as calling for a shift in the current research agenda. Instead of focusing predominantly on correlates and determinants of BIM, more valuable insights might be gained by allocating the available resources to studies on IB, its boundary conditions, and its underlying mental processes. BIM may still play a role in this endeavor by providing insights into the determinants of IB. However, the emphasis in such research would be different from the one in extant research on BIM, in that the primary focus is on IB as a phenomenon that needs to be understood, and BIM may serve as a tool to understand IB. Keeping the extant emphasis on BIM as the focal phenomenon would provide insights into IB only if BIM can be meaningfully equated with IB. The current analysis suggests that such an equation is highly problematic on both conceptual and empirical grounds.

Acknowledgements

We thank Galen Bodenhausen, Jan De Houwer, and Alex Madva for helpful comments on earlier versions of this article. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Funding

Preparation of this article was supported by National Science Foundation Grant BCS-1941440.

ORCID

Bertram Gawronski http://orcid.org/0000-0001-7938-3339
Alison Ledgerwood http://orcid.org/0000-0002-4535-6276
Paul W. Eastwick http://orcid.org/0000-0001-8512-8721

References

Ajzen, I. (1987). Attitudes, traits, and actions: Dispositional prediction of behavior in personality and social psychology. *Advances in Experimental Social Psychology*, *20*, 1–63.

Allison, S. T., & Kerr, N. L. (1994). Group correspondence biases and the provision of public goods. *Journal of Personality and Social Psychology*, *66*(4), 688–698. doi:10.1037/0022-3514.66.4.688

Allison, S. T., & Messick, D. M. (1985). The group attribution error. *Journal of Experimental Social Psychology*, *21*(6), 563–579. doi:10.1016/0022-1031(85)90025-3

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.

Banaji, M. R., & Hardin, C. D. (1996). Automatic stereotyping. *Psychological Science*, *7*(3), 136–141. doi:10.1111/j.1467-9280.1996.tb00346.x

Banse, R., Gawronski, B., Rebetez, C., Gutt, H., & Morton, J. B. (2010). The development of spontaneous gender stereotyping in childhood: Relations to stereotype knowledge and stereotype flexibility. *Developmental Science*, *13*(2), 298–306.

Bar-Anan, Y., Nosek, B. A., & Vianello, M. (2009). The sorting paired features task: A measure of association strengths. *Experimental Psychology*, *56*(5), 329–343.

Bargh, J. A. (1994). The four horsemen of automaticity: Awareness, intention, efficiency, and control in social cognition. In R. S. Wyer & T. K. Srull (Eds.), *Handbook of social cognition* (pp. 1–40). Hillsdale, NJ: Erlbaum.

Barnes-Holmes, D., Barnes-Holmes, Y., Stewart, I., & Boles, S. (2010). A sketch of the Implicit Relational Assessment Procedure (IRAP) and the Relational Elaboration and Coherence (REC) model. *The Psychological Record*, *60*(3), 527–542. doi:10.1007/BF03395726

Bijlstra, G., Holland, R. W., Dotsch, R., Hugenberg, K., & Wigboldus, D. H. J. (2014). Stereotype associations and emotion recognition. *Personality & Social Psychology Bulletin*, *40*(5), 567–577.

Bodenhausen, G. V. (1988). Stereotypic biases in social decision making and memory: Testing process models of stereotype use. *Journal of Personality and Social Psychology*, *55*(5), 726–737. doi:10.1037/0022-3514.55.5.726

Bonam, C. M., Nair Das, V., Coleman, B. R., & Salter, P. (2019). Ignoring history, denying racism: Mounting evidence for the Marley hypothesis and epistemologies of ignorance. *Social Psychological and Personality Science*, *10*(2), 257–265. doi:10.1177/1948550617751583

Brendl, C. M., Markman, A. B., & Messner, C. (2005). Indirectly measuring evaluations of several attitude objects in relation to a neutral reference point. *Journal of Experimental Social Psychology*, *41*(4), 346–368. doi:10.1016/j.jesp.2004.07.006

Brownstein, M., Madva, A., & Gawronski, B. (2019). What do implicit measures measure? *Wiley Interdisciplinary Reviews: Cognitive Science*, *10*(5), e1501.

Cameron, C. D., Brown-Iannuzzi, J., & Payne, B. K. (2012). Sequential priming measures of implicit social cognition: A meta-analysis of associations with behaviors and explicit attitudes. *Personality and Social Psychology Review : An Official Journal of the Society for Personality and Social Psychology, Inc*, *16*(4), 330–350.

Carter, E. R., Onyeador, I. N., & Lewis, N. A. Jr. (2020). Developing & delivering effective anti-bias training: Challenges & recommendations. *Behavioral Science & Policy*, *6*(1), 57–70. doi:10.1353/bsp.2020.0005

Cervone, D., Shadel, W. D., & Jencius, S. (2001). Social-cognitive theory of personality assessment. *Personality and Social Psychology Review*, *5*(1), 33–51. doi:10.1207/S15327957PSPR0501_3

Chen, M., & Bargh, J. A. (1999). Consequences of automatic evaluation: Immediate behavioral predispositions to approach or avoid the stimulus. *Personality and Social Psychology Bulletin*, *25*(2), 215–224. doi:10.1177/0146167299025002007

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

Corneille, O., & Hütter, M. (2020). Implicit? What do you mean? A comprehensive review of the delusive implicitness construct in attitude research. *Personality and Social Psychology Review*, *24*(3), 212–232. doi:10.1177/1088868320911325

Corneille, O., & Stahl, C. (2019). Associative attitude learning: A closer look at evidence and how it relates to attitude model. *Personality and Social Psychology Review*, *23*(2), 161–189. doi:10.1177/1088868318763261

Cummins, J., & De Houwer, J. (2019). An inkblot for beliefs: The Truth Misattribution Procedure. *PLoS One*, *14*(6), e0218661. doi:10.1371/journal.pone.0218661

Cunningham, W. A., Preacher, K. J., & Banaji, M. R. (2001). Implicit attitude measurement: Consistency, stability, and convergent validity. *Psychological Science*, *12*(2), 163–170. doi:10.1111/1467-9280.00328

Darley, J. M., & Gross, P. H. (1983). A hypothesis-confirming bias in labeling effects. *Journal of Personality and Social Psychology*, *44*(1), 20–33. doi:10.1037/0022-3514.44.1.20

Daumeyer, N. M., Onyeador, I. N., Brown, X., & Richeson, J. A. (2019). Consequences of attributing discrimination to implicit vs. explicit bias. *Journal of Experimental Social Psychology*, *84*, 103812. doi:10.1016/j.jesp.2019.04.010

Daumeyer, N. M., Rucker, J. M., & Richeson, J. A. (2017). Thinking structurally about implicit bias: Some peril, lots of promise. *Psychological Inquiry*, *28*(4), 258–261. doi:10.1080/1047840X.2017.1373556

De Houwer, J. (2003). The extrinsic affective Simon task. *Experimental Psychology*, *50*(2), 77–85. doi:10.1026//1618-3169.50.2.77

De Houwer, J. (2019). Implicit bias is behavior: A functional-cognitive perspective on implicit bias. *Perspectives on Psychological Science : A Journal of the Association for Psychological Science*, *14*(5), 835–840. doi:10.1177/1745691619855638

De Houwer, J., & De Bruycker, E. (2007). The identification-EAST as a valid measure of implicit attitudes toward alcohol-related stimuli. *Journal of Behavior Therapy and Experimental Psychiatry*, *38*(2), 133–143. doi:10.1016/j.jbtep.2006.10.004

De Houwer, J., Gawronski, B., & Barnes-Holmes, D. (2013). A functional-cognitive framework for attitude research. *European Review of Social Psychology*, *24*(1), 252–287. doi:10.1080/10463283.2014.892320

De Houwer, J., Heider, N., Spruyt, A., Roets, A., & Hughes, S. (2015). The relational responding task: Toward a new implicit measure of beliefs. *Frontiers in Psychology*, *6*, 319. doi:10.3389/fpsyg.2015.00319

De Houwer, J., Teige-Mocigemba, S., Spruyt, A., & Moors, A. (2009). Implicit measures: A normative analysis and review. *Psychological Bulletin*, *135*(3), 347–368.

De Houwer, J., Van Dessel, P., & Moran, T. (2020). Attitudes beyond associations: On the role of propositional representations in stimulus evaluation. *Advances in Experimental Social Psychology*, *61*, 127–183.

Degner, J., & Wentura, D. (2008). The extrinsic affective Simon task as an instrument for indirect assessment of prejudice. *European Journal of Social Psychology*, *38*(6), 1033–1043. doi:10.1002/ejsp.536

Diekman, A. B., & Eagly, A. H. (2000). Stereotypes as dynamic constructs: Women and men of the past, present, and future. *Personality and Social Psychology Bulletin*, *26*(10), 1171–1188. doi:10.1177/0146167200262001

Dovidio, J. F., & Gaertner, S. L. (2004). Aversive racism. *Advances in Experimental Social Psychology*, *36*, 1–52.

Duncan, B. L. (1976). Differential perception and attribution of intergroup violence: Testing the lower limits of stereotyping of Blacks. *Journal of Personality and Social Psychology*, *34*(4), 590–598. doi:10.1037/0022-3514.34.4.590

Dunham, Y., Baron, A. S., & Banaji, M. R. (2008). The development of implicit intergroup cognition. *Trends in Cognitive Sciences*, *12*(7), 248–253.

Dunton, B. C., & Fazio, R. H. (1997). An individual difference measure of motivation to control prejudiced reactions. *Personality and Social Psychology Bulletin*, *23*(3), 316–326. doi:10.1177/0146167297233009

Edwards, J. (2016, June 27). Justice Dept. mandates 'implicit bias' training for agents, lawyers. *Reuters, United States Edition*. Retrieved from http://www.reuters.com/article/us-usa-justice-bias-exclusive-idUSKCN0ZD251

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. H. (2007). Attitudes as object-evaluation associations of varying strength. *Social Cognition*, *25*(5), 603–637. doi:10.1521/soco.2007.25.5.603

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*(6), 1013–1027.

Fields, K. E., & Fields, B. J. (2014). *Racecraft: The soul of inequality in American life*. New York: Verso Trade.

Fleeson, W., & Jayawickreme, E. (2021). Whole traits: Revealing the social-cognitive mechanisms constituting personality's central variable. *Advances in Experimental Social Psychology*, *63*, 69–128.

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

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

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

Gawronski, B. (2004). Theory-based bias correction in dispositional inference: The fundamental attribution error is dead, long live the correspondence bias. *European Review of Social Psychology*, *15*(1), 183–217. doi:10.1080/10463280440000026

Gawronski, B. (2019). Six lessons for a cogent science of implicit bias and its criticism. *Perspectives on Psychological Science : A Journal of the Association for Psychological Science*, *14*(4), 574–595. doi:10.1177/1745691619826015

Gawronski, B., & Bodenhausen, G. V. (2006). Associative and propositional processes in evaluation: An integrative review of implicit and explicit attitude change. *Psychological Bulletin*, *132*(5), 692–731. doi:10.1037/0033-2909.132.5.692

Gawronski, B., & Bodenhausen, G. V. (2012). Self-insight from a dual-process perspective. In S. Vazire & T. D. Wilson (Eds.), *Handbook of self-knowledge* (pp. 22–38). New York: Guilford Press.

Gawronski, B., & Bodenhausen, G. V. (2015a). Social-cognitive theories. In B. Gawronski, & G. V. Bodenhausen (Eds.), *Theory and explanation in social psychology* (pp. 65–83). New York: Guilford Press.

Gawronski, B., & Bodenhausen, G. V. (2015b). Theory evaluation. In B. Gawronski, & G. V. Bodenhausen (Eds.), *Theory and explanation in social psychology* (pp. 3–23). New York: Guilford Press.

Gawronski, B., & Bodenhausen, G. V. (2017). Beyond persons and situations: An interactionist approach to understanding implicit bias. *Psychological Inquiry*, *28*(4), 268–272. doi:10.1080/1047840X.2017.1373546

Gawronski, B., & Brannon, S. M. (2019). Attitudes and the implicit-explicit dualism. In D. Albarracín & B. T. Johnson (Eds.), *The handbook of attitudes. Volume 1: Basic principles* (2nd ed., pp. 158–196). New York, NY: Routledge.

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 ed., pp. 283–310). New York: Cambridge University Press.

Gawronski, B., De Houwer, J., & Sherman, J. W. (2020). Twenty-five years of research using implicit measures. *Social Cognition*, *38*(Suppl), s1–s25. doi:10.1521/soco.2020.38.supp.s1

Gawronski, B., Geschke, D., & Banse, R. (2003). Implicit bias in impression formation: Associations influence the construal of individuating information. *European Journal of Social Psychology*, *33*(5), 573–589. doi:10.1002/ejsp.166

Gawronski, B., Hofmann, W., & Wilbur, C. J. (2006). Are "implicit "attitudes unconscious? *Consciousness and Cognition*, *15*(3), 485–499.

Gawronski, B., LeBel, E. P., & Peters, K. R. (2007). What do implicit measures tell us? Scrutinizing the validity of three common assumptions. *Perspectives on Psychological Science : A Journal of the Association for Psychological Science*, *2*(2), 181–193.

Gawronski, B., Ledgerwood, A., & Eastwick, P. W. (2020). Implicit bias and anti-discrimination policy. *Policy Insights from the Behavioral and Brain Sciences*, *7*(2), 99–106. doi:10.1177/2372732220939128

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

Gilbert, D. T., & Jones, E. E. (1986). Perceiver-induced constraint: Interpretations of self-generated reality. *Journal of Personality and Social Psychology*, *50*(2), 269–280. doi:10.1037/0022-3514.50.2.269

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review*, *102*(1), 4–27.

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

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*(4), 553–561.

Greenwald, A. G., & Lai, C. K. (2020). Implicit social cognition. *Annual Review of Psychology*, *71*, 419–445.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. K. L. (1998). Measuring individual differences in implicit cognition: The Implicit Association Test. *Journal of Personality and Social Psychology*, *74*(6), 1464–1480. doi:10.1037/0022-3514.74.6.1464

Greenwald, A. G., Poehlman, T. A., Uhlmann, E., & Banaji, M. R. (2009). Understanding and using the Implicit Association Test: III. Meta-analysis of predictive validity. *Journal of Personality and Social Psychology*, *97*(1), 17–41.

Griggs, B. (2018, December 28). Living while black: Here are all the routine activities for which police were called on African-Americans this year. *CNN*. Retrieved from https://www.cnn.com/2018/12/20/us/living-while-black-police-calls-trnd/index.html

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*(2), 70–79. doi:10.1027/1614-0001.29.2.70

Hahn, A., & Gawronski, B. (2019). Facing one's implicit biases: From awareness to acknowledgment. *Journal of Personality and Social Psychology*, *116*(5), 769–794.

Hahn, A., & Goedderz, A. (2020). Trait-unconsciousness, state-unconsciousness, preconsciousness, and social miscalibration in the context of implicit evaluations. *Social Cognition*, *38*(Suppl), s115–s134. doi:10.1521/soco.2020.38.supp.s115

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

Hailey, C. A. (2020). *Choosing schools, choosing safety: The role of school safety in school choice*. (Unpublished dissertation). New York University, New York, NY.

Hehman, E., Calanchini, J., Flake, J. K., & Leitner, J. B. (2019). Establishing construct validity evidence for regional measures of explicit and implicit racial bias. *Journal of Experimental Psychology: General*, *148*(6), 1022–1040. doi:10.1037/xge0000623

Helms, J. E. (1997). Toward a model of White racial identity development. In P. G. Altbach, K. Arnold, & I. C. King (Eds.), *College student development and academic life: Psychological, intellectual, social and moral issues* (pp. 49–66). New York: Routledge.

Hodson, G., Dovidio, J. F., & Gaertner, S. L. (2002). Processes in racial discrimination: Differential weighting of conflicting information. *Personality and Social Psychology Bulletin*, *28*(4), 460–471. doi:10.1177/0146167202287004

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 & Social Psychology Bulletin*, *31*(10), 1369–1385.

Honey, M. (1984). *Creating Rosie the Riveter: Class, gender, and propaganda during World War II*. Amherst: University of Massachusetts Press.

Hooks, B. (2003). *Teaching community: A pedagogy of hope*. New York: Taylor & Francis.

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*(1), 125–145. doi:10.1111/bjso.12168

Hugenberg, K., & Bodenhausen, G. V. (2003). Facing prejudice: Implicit prejudice and the perception of facial threat. *Psychological Science*, *14*(6), 640–643.

Hughes, S., Cummins, J., & Hussey, I. (in press). Effects on the Affect Misattribution Procedure are strongly moderated by awareness. *PsyArXiv*. doi:10.31234/osf.io/d5zn8

Humphrey, R. (1985). How work roles influence perception: Structural-cognitive processes and organizational behavior. *American Sociological Review*, *50*(2), 242–252. doi:10.2307/2095412

Hutchings, P. B., & Haddock, G. (2008). Looking black in anger: The role of implicit prejudice in the categorization and perceived emotional intensity of racially ambiguous faces. *Journal of Experimental Social Psychology*, *44*(5), 1418–1420. doi:10.1016/j.jesp.2008.05.002

Jones, J. M., & Carter, R. T. (1996). Racism and White racial identity: Merging realities. In B. P. Bowser, & R. G. Hunt (Eds.), *Impacts of racism on White Americans* (pp. 1–23). Thousand Oaks, CA: Sage.

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), 16–32.

Klauer, K. C., Voss, A. & Stahl, C. (Eds.). (2011). *Cognitive methods in social psychology*. New York: Guilford Press.

Krickel, B. (2018). Are the states underlying implicit biases unconscious? - A neo-Freudian answer. *Philosophical Psychology*, *31*(7), 1007–1026. doi:10.1080/09515089.2018.1470323

Kruglanski, A. W. (1989). The psychology of being "right": The problem of accuracy in social perception and cognition. *Psychological Bulletin*, *106*(3), 395–409. doi:10.1037/0033-2909.106.3.395

Kruglanski, A. W., & Ajzen, I. (1983). Bias and error in human judgment. *European Journal of Social Psychology*, *13*(1), 1–44. doi:10.1002/ejsp.2420130102

Kunda, Z., & Sherman-Williams, B. (1993). Stereotypes and the construal of individuating information. *Personality and Social Psychology Bulletin*, *19*(1), 90–99. doi:10.1177/0146167293191010

Kurdi, B., & Dunham, Y. (2020). Propositional accounts of implicit evaluation: Taking stock and looking ahead. *Social Cognition*, *38*(Suppl), s42–s67. doi:10.1521/soco.2020.38.supp.s42

Kurdi, B., Seitchik, A. E., Axt, J. R., Carroll, T. J., Karapetyan, A., Kaushik, N., … Banaji, M. R. (2019). Relationship between the Implicit Association Test and intergroup behavior: A meta-analysis. *The American Psychologist*, *74*(5), 569–586.

Lai, C. K., Marini, M., Lehr, S. A., Cerruti, C., Shin, J.-E L., Joy-Gaba, J. A., … Nosek, B. A. (2014). A comparative investigation of 18 interventions to reduce implicit racial preferences. *Journal of Experimental Psychology: General*, *143*(4), 1765–1785. doi:10.1037/a0036260

Lai, C. K., & Wilson, M. E. (2020). Measuring implicit intergroup biases. *Social and Personality Psychology Compass*, *15*, e12573.

Ledgerwood, A., Eastwick, P. W., & Gawronski, B. (2020). Experiences of liking versus ideas about liking. *The Behavioral and Brain Sciences*, *43*, e136.

Ledgerwood, A., Hudson, S. T. J., Lewis, N. A., Jr., Maddox, K. B., Pickett, C. L., Remedios, J. D., … Wilkins, C. L. (2021). *The pandemic as a portal: Reimagining psychological science as truly open and inclusive*. Unpublished manuscript. doi:10.31234/osf.io/gdzue

MacDonald, H. (2017, October 9). The false "science" of implicit bias. *Wall Street Journal*. Retrieved from https://www.wsj.com/articles/the-false-science-of-implicit-bias-1507590908

Mackie, D. M., & Allison, S. T. (1987). Group attribution errors and the illusion of group attitude change. *Journal of Experimental Social Psychology*, *23*(6), 460–480. doi:10.1016/0022-1031(87)90016-3

Mallon, R. (2021). Racial attitudes, accumulation mechanisms, and disparities. *Review of Philosophy and Psychology*, *12*, 953–975.

McConnell, A. R., & Leibold, J. M. (2001). Relations among the Implicit Association Test, discriminatory behavior, and explicit measures of racial attitudes. *Journal of Experimental Social Psychology*, *37*(5), 435–442. doi:10.1006/jesp.2000.1470

Monteith, M. J., Voils, C. I., & Ashburn-Nardo, L. (2001). Taking a look underground: Detecting, interpreting, and reacting to implicit racial biases. *Social Cognition*, *19*(4), 395–417. doi:10.1521/soco.19.4.395.20759

Moors, A. (2016). Automaticity: Componential, causal, and mechanistic explanations. *Annual Review of Psychology*, *67*, 263–287. doi:10.1146/annurev-psych-122414-033550

Mueller, J. C. (2017). Producing colorblindness: Everyday mechanisms of white ignorance. *Social Problems*, *64*(2), 238–332. doi:10.1093/socpro/spx012

Neville, H. A., Awad, G. H., Brooks, J. E., Flores, M. P., & Bluemel, J. (2013). Color-blind racial ideology: Theory, training, and measurement implications in psychology. *The American Psychologist*, *68*(6), 455–466.

Newell, B. R., & Shanks, D. R. (2014). Unconscious influences on decision-making: A critical review. *The Behavioral and Brain Sciences*, *37*(1), 1–19.

Norton, M. I., Vandello, J. A., & Darley, J. M. (2004). Casuistry and social category bias. *Journal of Personality and Social Psychology*, *87*(6), 817–831.

Nosek, B. A., & Banaji, M. R. (2001). The go/no-go association task. *Social Cognition*, *19*(6), 625–666. doi:10.1521/soco.19.6.625.20886

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

Onyeador, I. N., Hudson, S. K. T., & Lewis, N. A. Jr. (2021). Moving beyond implicit bias training: Policy insights for increasing organizational diversity. *Policy Insights from the Behavioral and Brain Sciences*, *8*(1), 19–26. doi:10.1177/2372732220983840

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*(2), 171–192.

Payne, B. K., Brown-Iannuzzi, J., Burkley, M., Arbuckle, N. L., Cooley, E., Cameron, C. D., & Lundberg, K. B. (2013). Intention invention and the Affect Misattribution Procedure: Reply to Bar-Anan and Nosek (2012). *Personality & Social Psychology Bulletin*, *39*(3), 375–386. doi:10.1177/0146167212475225

Payne, B. K., Cheng, S. 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, B. K., & Correll, J. (2020). Race, weapons, and the perception of threat. *Advances in Experimental Social Psychology*, *62*, 1–50.

Payne, B. K., Niemi, L., & Doris, J. M. (2018). How to think about "implicit bias." *Scientific American*. Retrieved from https://www.scientificamerican.com/article/how-to-think-about-implicit-bias/

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

Penke, L., Eichstaedt, J., & Asendorpf, J. B. (2006). Single Attribute Implicit Association Tests (SA-IAT) for the assessment of unipolar constructs: The case of sociosexuality. *Experimental Psychology*, *53*(4), 283–291.

Plant, E. A., & Devine, P. G. (1998). Internal and external motivation to respond without prejudice. *Journal of Personality and Social Psychology*, *75*(3), 811–832. doi:10.1037/0022-3514.75.3.811

Roediger, H. L. (1990). Implicit memory: Retention without remembering. *The American Psychologist*, *45*(9), 1043–1056. doi:10.1037/0003-066X.45.9.1043

Ross, L., & Ward, A. (1996). Naive realism in everyday life: Implications for social conflict and misunderstanding. In E. S. Reed,

E. Turiel, & T. Brown (Eds.), *The Jean Piaget symposium series: Values and knowledge* (pp. 103–135). Mahwah, NJ: Lawrence Erlbaum.

Ross, L. D., Amabile, T. M., & Steinmetz, J. L. (1977). Social roles, social control, and biases in social-perception processes. *Journal of Personality and Social Psychology*, *35*(7), 485–494. doi:10.1037/0022-3514.35.7.485

Rothermund, K., Teige-Mocigemba, S., Gast, A., & Wentura, D. (2009). Minimizing the influence of recoding in the IAT: The Recoding-Free Implicit Association Test (IAT-RF). *Quarterly Journal of Experimental Psychology*, *62*(1), 84–98. doi:10.1080/17470210701822975

Rothstein, R. (2017). The color of law: A forgotten history of how our government segregated America. New York: Liveright Publishing Corporation.

Sagar, H. A., & Schofield, J. W. (1980). Racial and behavioral cues in black and white children's perceptions of ambiguously aggressive acts. *Journal of Personality and Social Psychology*, *39*(4), 590–598.

Salter, P. S., Adams, G., & Perez, M. J. (2018). Racism in the structure of everyday worlds: A cultural-psychological perspective. *Current Directions in Psychological Science*, *27*(3), 150–155. doi:10.1177/0963721417724239

Schacter, D. L. (1987). Implicit memory: History and current status. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *13*, 501–518.

Schimmack, U. (2021). The Implicit Association Test: A method in search of a construct. *Perspectives on Psychological Science*, *16*(2), 396–414. doi:10.1177/1745691619863798

Schnabel, K., Banse, R., & Asendorpf, J. B. (2006). Employing automatic approach and avoidance tendencies for the assessment of implicit personality self-concept: The Implicit Association Procedure (IAP). *Experimental Psychology*, *53*(1), 69–76. doi:10.1027/1618-3169.53.1.69

Shanks, D. R., & St. John, M. F. (1994). Characteristics of dissociable human learning systems. *Behavioral and Brain Sciences*, *17*(3), 367–447. doi:10.1017/S0140525X00035032

Sriram, N., & Greenwald, A. G. (2009). The Brief Implicit Association Test. *Experimental Psychology*, *56*(4), 283–294.

Strack, F., & Deutsch, R. (2004). Reflective and impulsive determinants of social behavior. *Personality and Social Psychology Review: An Official Journal of the Society for Personality and Social Psychology, Inc*, *8*(3), 220–247.

Strack, F., & Hannover, B. (1996). Awareness of the influence as a precondition for implementing correctional goals. In P. M. Gollwitzer & J. A. Bargh (Eds.), *The psychology of action: Linking cognition and motivation to behavior* (pp. 579–596). New York: Guilford Press.

Sweldens, S., Corneille, O., & Yzerbyt, V. (2014). The role of awareness of attitude formation via evaluative conditioning. *Personality and Social Psychology Review*, *18*(2), 187–209. doi:10.1177/1088868314527832

Teige-Mocigemba, S., Klauer, K. C., & Rothermund, K. (2008). Minimizing method-specific variance in the IAT: The Single Block IAT. *European Journal of Psychological Assessment*, *24*(4), 237–245. doi:10.1027/1015-5759.24.4.237

Trope, Y. (1986). Identification and inferential processes in dispositional attribution. *Psychological Review*, *93*(3), 239–257. doi:10.1037/0033-295X.93.3.239

UCnet (2021). *UC managing implicit bias series*. Retrieved from https://ucnet.universityofcalifornia.edu/working-at-uc/your-career/talent-management/professional-development/managing-implicit-bias.html

Uhlmann, E. L., & Cohen, G. L. (2005). Constructed criteria: Redefining merit to justify discrimination. *Psychological Science*, *16*(6), 474–480. doi:10.1111/j.0956-7976.2005.01559.x

Webb, E. J., Campbell, D. T., Schwartz, R. D., & Sechrist, L. (1966). *Unobtrusive measures: Nonreactive research in the social sciences*. Chicago, IL: Rand McNally.

Wegener, D. T., & Petty, R. E. (1997). The flexible correction model: The role of naive theories of bias in bias correction. *Advances in Experimental Social Psychology*, *29*, 141–208.

Wellman, J. D., Wilkins, C. L., Newell, E. E., & Stewart, D. K. (2019). Conflicting motivations: Understanding how low-status group members respond to ingroup discrimination claimants. *Personality & Social Psychology Bulletin*, *45*(8), 1170–1183.

Wilson, T. D., & Brekke, N. (1994). Mental contamination and mental correction: Unwanted influences on judgments and evaluations. *Psychological Bulletin*, *116*(1), 117–142.

Wittenbrink, B., Judd, C. M., & Park, B. (1997). Evidence for racial prejudice at the implicit level and its relationships with questionnaire measures. *Journal of Personality and Social Psychology*, *72*(2), 262–274. doi:10.1037/0022-3514.72.2.262

Worth, L. T., Allison, S. T., & Messick, D. M. (1987). Impact of a group decision on perception of one's own and others' attitudes. *Journal of Personality and Social Psychology*, *53*(4), 673–682. doi:10.1037/0022-3514.53.4.673