Six Lessons for a Cogent Science of Implicit Bias and Its Criticism

Bertram Gawronski
University of Texas at Austin

Skepticism about the explanatory value of *implicit bias* in understanding social discrimination has grown considerably. The current article argues that both the dominant narrative about implicit bias as well as extant criticism are based on a selective focus on particular findings that fails to consider the broader literature on attitudes and implicit measures. To provide a basis to move forward, the current article discusses six lessons for a cogent science of implicit bias: (1) There is no evidence that people are unaware of the mental contents underlying their implicit biases. (2) Conceptual correspondence is essential for interpretations of dissociations between implicit and explicit bias. (3) There is no basis to expect strong unconditional relations between implicit bias and behavior. (4) Implicit bias is less (not more) stable over time than explicit bias. (5) Context matters fundamentally for the outcomes obtained with implicit bias measures. (6) Implicit measurement scores do not provide process-pure reflections of bias. The six lessons provide guidance for research that aims to provide more compelling evidence for the properties of implicit bias. At the same time, they suggest that extant criticism does not justify the conclusion that implicit bias is irrelevant for the understanding of social discrimination.

Keywords: attitudes; consciousness; context effects; implicit bias; measurement

Like many other high-profile phenomena in social psychology (e.g., Friese, Loschelder, Gieseler, Frankenach, & Inzlicht, in press; Molden, 2014; Wagenmakers et al., 2016), research on implicit bias has become the target of increased scrutiny. Although critics have expressed concerns about the meaning and significance of the *implicit bias* construct for more than a decade (e.g., Arkes & Tetlock, 2004; Fiedler, Messner, & Bluemke, 2006), skeptic views have received significantly more attention over the last few years. In fact, the growing skepticism has become so pervasive that even early proponents have started to question the explanatory value of implicit bias (e.g., Forscher, Mitamura, Dix, Cox, & Devine, 2017), with some critics dismissing the construct as entirely irrelevant for the psychological understanding of social discrimination (e.g., Blanton & Jaccard, 2017; Mitchell, 2018). Similar shifts can be found in the coverage of implicit bias research in the popular media. Although references to implicit bias in the public discourse about social discrimination are at an all-time high (e.g., Baker, 2018; Clinton, 2016; Whitten, 2018), criticism of implicit bias research is receiving much more attention, which is reflected in critical headlines such as *Can We Really Measure Implicit Bias? Maybe Not* (Bartlett, 2017) or *The False 'Science' of Implicit Bias* (MacDonald, 2017).

In the current article, I argue that both the mainstream narrative about implicit bias as well as extant criticism of implicit bias research have failed to consider key insights in the broader literature on attitudes and implicit measures (see Albarracín & Johnson, 2019; Gawronski & Payne, 2010). Although these insights pose other unacknowledged challenges to the mainstream narrative about implicit bias, they suggest that at least some of the dominant criticism is based on a selective focus on particular findings that ignores key insights in the broader literature. At the same time, an expanded focus that includes the broader literature on attitudes and implicit measures suggests that the meaning of numerous findings is ambiguous and that, therefore, many dominant questions about implicit bias remain unanswered.

To provide common ground and a basis to move forward, the current article discusses six lessons for an empirically, theoretically, and methodologically informed science of implicit bias and critical debates about the range and limits of the construct in understanding the psychological underpinnings of social discrimination.[1] Together, the six lessons suggest that research on implicit bias would benefit from considering the broader literature on implicit measures as well as historical debates in research on attitudes. At the same time, they suggest that the criticisms raised against research on implicit bias do not justify the inference that the construct is entirely irrelevant for the psychological understanding of social discrimination. The main conclusion is that future research adhering to the normative implications of the six lessons is essential for a more nuanced understanding of implicit bias, its psychological characteristics, and its potential contribution to social discrimination.

[1] The six lessons are not meant to be exhaustive in the sense that they address all criticisms that have been raised against particular measurement instruments in implicit bias research. Instead, they are meant to provide a common basis for future research on implicit bias irrespective of the employed instruments. Because the shortcomings of one instrument can often be compensated by the strengths of another instrument (and vice versa), it seems possible to rule out instrument-specific criticism by replicating a given finding with different instruments. To the extent that a finding replicates for multiple instruments with unique strengths and weaknesses, greater confidence can be gained regarding the reliability and theoretical meaning of the obtained effect (see Lesson 6).

Lesson 1: There is no evidence that people are unaware of the mental contents underlying their implicit biases.

Historically, the development of implicit measures can be traced back to two independent lines of research with distinct conceptual roots (Payne & Gawronski, 2010). One the one hand, the development of the evaluative priming task (EPT; Fazio, Jackson, Dunton, & Williams, 1995) was based on the idea that attitudes, conceptualized as object-evaluation associations in memory, can be activated automatically to the extent that the association between the attitude object and its stored summary evaluation is sufficiently strong (see Fazio, 2007). On the other hand, the development of the implicit association test (IAT; Greenwald, McGhee, & Schwartz, 1998) was inspired by research on implicit memory, suggesting that past experiences can influence responses in the absence of explicit memory for the relevant experiences (see Greenwald & Banaji, 1995). Although the EPT and the IAT are just two among more than a dozen implicit measures that are available to date (for a review, see Gawronski & De Houwer, 2014), most research on implicit bias has relied on either one or the other conceptualization. Whereas research guided by the conceptual roots of the EPT tends to emphasize the unintentionality, efficiency, and uncontrollability of attitude activation without any claims of unawareness (see Bargh, 1994), research guided by the conceptual roots of the IAT emphasizes the idea that people are unaware of the mental contents underlying their responses on implicit measures.

Claims of unawareness are often based on the methodological truism that implicit measures, in contrast to explicit measures, do not require that participants are aware of the to-be-measured mental contents (Greenwald & Banaji, 1995). Whereas accurate self-reports on explicit measures presuppose that participants are aware of the to-be-measured mental contents, implicit measures do not require awareness, because participants are not directly asked about them. Instead, mental contents are inferred from participants' performance (e.g., speed and/or accuracy) on experimental paradigms based on sequential priming or response interference (for a review, see Gawronski & De Houwer, 2014). Based on this methodological difference, it is often assumed that explicit measures capture conscious biases, whereas implicit measures capture unconscious biases (e.g., Cunningham, Nezlek, & Banaji, 2004; Rudman, Greenwald, Mellott, & Schwartz, 1999).

Because implicit measures do not require awareness of the to-be-measured mental contents, they certainly have the potential to capture unconscious mental contents that evade assessment via explicit measures. However, this possibility does not imply that people are unaware of the mental contents underlying their responses on implicit measures. Any such claim is an empirical hypothesis that has to be evaluated based on relevant evidence (De Houwer, Teige-Mocigemba, Spruyt, & Moors, 2009). Indeed, a closer look at the available evidence raises serious doubts about the veracity of this hypothesis (for reviews, see Gawronski, Hofmann, & Wilbur, 2006; Gawronski, LeBel, & Peters, 2007).

A common argument in favor of the unawareness hypothesis is that correlations between implicit and explicit measures tend to be rather low (for meta-analyses, see Cameron, Brown-Iannuzzi, & Payne, 2012, Hofmann, Gawronski, Gschwendner, Le, & Schmitt, 2005). The theoretical idea underlying this argument is that unawareness of the mental contents captured by implicit measures makes it impossible to verbally report these contents on an explicit measure, which should lead to low correlations between implicit and explicit measures. Of course, correlations between the two kinds of measures can be expected to be low if people are unaware of the mental contents captured by implicit measures. However, correlations between implicit and explicit measures can be low for various other reasons that have nothing to do with lack of awareness (for a review, see Hofmann, Gschwendner, Nosek, & Schmitt, 2005). In the area of intergroup bias, for example, several studies found that correlations between implicit and explicit measures are significantly higher among participants with low motivation to control prejudiced reactions compared to participants with high motivation to control prejudiced reactions (e.g., Degner & Wentura, 2008; Dunton & Fazio, 1997; Gawronski, Geschke, & Banse, 2003; Payne, Cheng, Govorun, & Stewart, 2005). Although it might be possible to reconcile this finding with the unawareness hypothesis in a post-hoc fashion, it is predicted a priori by extant theories suggesting that verbal reports of activated mental contents depend on the motivation and the opportunity to control their expression (Fazio, 2007; Fazio & Towles-Schwen, 1999). According to this view, correlations between implicit and explicit measures should be low when participants have both the motivation and the opportunity to control the expression of activated mental contents. In contrast, correlations between the two kinds of measures should be high when participants lack either the motivation or the opportunity to control the expression of activated mental contents.

More direct evidence against the unawareness hypothesis comes from research by Hahn, Judd, Hirsh, and Blair (2014) who investigated whether participants are able to predict their scores on implicit measures (see also Hahn & Gawronski, in press). In a series of studies, participants were asked to predict their scores on multiple IATs capturing attitudes toward different social groups and then completed the same IATs.

Counter to the widespread assumption that participants are unaware of the mental contents captured by the IAT, participants were able to predict the pattern of their IAT scores with a high degree of accuracy (i.e., median correlations between predicted and actual patterns of IAT scores around .65). Accuracy in the prediction of IAT scores was high regardless of participants' prior experience with the IAT, regardless of how much information participants received about the IAT, and regardless of whether the IAT was described as a measure of "true beliefs" or "cultural associations." Moreover, predicted and actual IAT scores were highly correlated although self-reported evaluations on explicit measures showed the same low correlations with IAT scores that are typically observed in this area (see Cameron et al., 2012; Hofmann, Gawronski, et al., 2005). These findings pose a challenge to the hypothesis that people are unaware of the mental contents captured by implicit measures.[2]

Hahn et al.'s (2014) studies also debunk another common argument in favor of the unawareness hypothesis. Many visitors of the Project Implicit website are quite surprised when they are informed about their IAT performance (Howell, Gaither, & Ratliff, 2015; Howell & Ratliff, 2017), suggesting that the feedback they receive on their level of implicit bias deviates from their prior assumptions about their personal level of implicit bias. Such surprise reactions have been interpreted as evidence for the unawareness hypothesis, in that people should not be surprised about their IAT feedback if they were aware of their personal level of implicit bias (e.g., Banaji, 2011; Krickel, 2018). However, surprise reactions can also occur when the metric used to convert participants' numeric IAT scores into verbal feedback (e.g., *strong preference for Whites compared to Blacks*) deviates from participants' naïve metric in labeling their personal level of implicit bias. The findings by Hahn et al. (2014) are consistent with this argument, showing that, although participants are highly accurate in predicting their patterns of IAT scores, their naïve metric to label different levels of implicit bias "stretches" 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 arbitrary in the sense that there is no objective basis to treat one metric as "correct" and the other one as "incorrect" (Kruglanski, 1989), interpretations of surprise reactions as evidence for the unawareness hypothesis seem premature and empirically questionable.[3]

Although the currently available evidence poses a challenge to the hypothesis that people are unaware of the mental contents underlying their responses on implicit measures (e.g., Hahn et al., 2014), people may still be unaware of either the origin or the effects of these mental contents (or both). For example, based on a review of the available evidence, Gawronski et al. (2006) concluded that people are sometimes unaware of the origin of the mental contents underlying their responses on implicit measures. However, the same is true for the mental contents underlying responses on explicit measures, in that people are often unable to identify the causes of their self-reported preferences (for reviews, see Gawronski & Bodenhausen, 2012; Wilson, Dunn, Kraft, & Lisle, 1989). That is, people often know very well how much they like or dislike a given object and they are perfectly able to report their subjective evaluation on a self-report measure, but they may not know why they like or dislike the object (as captured by the popular phrase *I like it, but I don't know why*). Thus, although people are sometimes unaware of the origin of the mental contents captured by implicit measures, lack of source awareness does not seem to be a feature that distinguishes mental contents captured by implicit measures from mental contents captured by explicit measures (see Gawronski et al., 2006).

A more promising candidate seems to be the impact of the mental contents captured by implicit measures. Based on their review of the available evidence, Gawronski et al. (2006) concluded that (1) the mental contents underlying implicit measures may influence judgments and behavior outside of awareness and (2) such unconscious influences may not occur for the mental contents captured by explicit measures (see Gawronski et al., 2006). In line with this conclusion,

[2] One anonymous reviewer noted that prediction accuracy might be lower when participants were asked to predict their level of racial bias instead of their relative preference for social groups. Although it seems possible that prediction accuracy is lower under these conditions, unawareness of one's mental contents implies a general inability to predict implicit bias levels and this inability should be independent of the question framing. Thus, the fact that participants were able to predict their IAT scores in Hahn et al.'s (2014) studies suggests that (1) people are aware of the mental contents underlying their implicit biases and (2) potentially lower prediction accuracy on questions that include morally charged language reflects unwillingness to verbally label these mental contents as instances of racial bias (rather than a general inability to report them).

[3] One anonymous reviewer wondered if surprise about the existence of implicit bias as phenomenon would provide evidence for the unawareness hypothesis. Conceptually, however, people's knowledge of the existence of implicit bias as a phenomenon is independent of people's awareness of the mental contents underlying their implicit biases. On the one hand, educating people about the existence of implicit bias as a phenomenon does not necessarily increase people's awareness of the mental contents underlying their own implicit biases. On the other hand, people may be perfectly aware of the mental contents underlying their implicit biases even when they have never heard about implicit bias as a phenomenon.

Gawronski et al. (2003) found that participants interpreted ambiguous behavior by an outgroup member more negatively compared to the same behavior by an ingroup member, and the relative size of this effect was positively related to participants' implicit intergroup bias on an IAT (see also Hugenberg & Bodenhausen, 2003). There was no relation between biased interpretations of ambiguous behavior and participants' explicit intergroup bias. Interestingly, the obtained relation between implicit intergroup bias and biased interpretations of ambiguous behavior was unaffected by participants' motivation to control prejudiced reactions. That is, higher levels of implicit intergroup bias were associated with greater bias in the interpretation of ambiguous behavior even when participants were highly motivated to control prejudiced reactions. Yet, motivation to control did moderate the relation between implicit and explicit intergroup bias, in that implicit and explicit bias were positively related only for participants with low motivation to control prejudiced reactions, but not for participants with high motivation to control prejudiced reactions (see Degner & Wentura, 2008; Dunton & Fazio, 1997; Payne et al., 2005). Drawing on extant theories of bias correction (Strack & Hannover, 1996; Wegener & Petty, 1997), Gawronski et al. (2003) interpreted these findings as evidence for the hypothesis that the mental contents captured by implicit measures influence the processing of ambiguous information outside of awareness, leading to biased interpretations of ambiguous behavior even when people are motivated to control prejudiced reactions.

Although Gawronski et al.'s (2003) findings are consistent with this conclusion, their study suffers from a number of methodological limitations, one being that type of bias measure (implicit vs. explicit) was confounded with the specific contents of the two measures (evaluative responses to faces in the implicit measure vs. agreement with statements about cultural differences and perceived group relations in the explicit measure). Thus, it is unclear whether the obtained results reflect (1) a genuine difference between implicit and explicit bias or (2) a spurious difference that was driven by the different contents of the two bias measures (see Lesson 2 for a more detailed discussion of this issue). These ambiguities undermine the possibility of drawing strong conclusions from Gawronski et al.'s (2003) findings. Moreover, although lack of impact awareness seems consistent with a broad range of findings in the implicit bias literature (e.g., observed relations between implicit bias scores and measures of seating distance and nonverbal behavior; see Dovidio, Kawakami, & Gaertner, 2002; Fazio et al. 1995), there are no other studies that have directly tested this hypothesis with appropriate designs and awareness measures. Thus, despite common claims regarding lack of impact awareness, compelling evidence for these claims is surprisingly scarce.[4]

Implications

Lesson 1 suggests that statements about unawareness should be treated as hypotheses that require empirical evidence (see De Houwer et al., 2009). Moreover, because implicit biases have multiple aspects that could be outside of awareness, it is essential to clearly specify which aspect is assumed to be outside of awareness (see Gawronski et al., 2006). Do claims about unawareness refer to (1) the mental contents underlying responses on implicit bias measures (*content awareness*), (2) the origin of the underlying mental contents (*source awareness*), or (3) effects of the underlying mental contents on judgments and behavior (*impact awareness*)? Because some aspects of unawareness may be common for both implicit and explicit bias (e.g., lack of source awareness), researchers should also specify whether unawareness of particular aspect is assumed to be a unique feature of implicit bias that distinguishes it from explicit bias and provide empirical evidence for these hypotheses. If it is not possible to provide such evidence, it would seem appropriate to refrain from making strong claims about unawareness or to explicitly describe such claims as speculative. In fact, counter to a widespread assumption in the literature, there is currently no evidence that people are unaware of the mental contents underlying their responses on implicit measures. If anything, the available evidence suggests that people are aware of the mental contents underlying implicit measures, which allows them to predict their implicit bias scores with a high degree of accuracy (Hahn et al., 2014). Of course, it is possible that future research will pose a challenge

[4] A related question is whether participants are aware of the effect of their mental contents on their responses underlying implicit measures. For example, although many participants notice differences in their performance on the two combined blocks of the IAT (Monteith, Voils, & Ashburn-Nado, 2001), it seems unlikely that participants notice the rather small reaction time differences on different kinds of trials in the EPT (Petty, Fazio, & Briñol, 2009). Empirical evidence for the latter idea would be consistent with the hypothesis that people can be unaware of the behavioral effects of the mental contents captured by implicit measures. However, unawareness of behavioral effects does not permit any conclusions regarding unawareness of the mental contents themselves (e.g., people being aware of the mental contents underlying their responses on the IAT, but not of the mental contents underlying their responses on the EPT). Ironically, such (flawed) conclusions would also be inconsistent with the conceptual roots of the IAT and EPT, given that the concept of *implicit memory* played a major role for the development of the IAT (Greenwald & Banaji, 1995), but has been explicitly rejected as a conceptual basis for the EPT (Fazio, 2007).

to this conclusion by (1) providing the kind of evidence for the content unawareness hypothesis that is currently lacking; (2) questioning the reliability of previous evidence against the content unawareness hypothesis; or (3) providing new evidence that reconciles previous findings with the content unawareness hypothesis. However, in the absence of such evidence, it would seem appropriate to refrain from making empirically unsubstantiated claims about lack of content awareness in the interpretation of empirical findings. The same conclusion applies to claims about lack of source awareness and lack of impact awareness, which should be tested with appropriate designs and reliable measures of awareness. At this point, the available evidence suggests that people can be unaware of the origin of their implicit biases, but the same is true of explicit biases. Moreover, there is preliminary evidence that implicit, but not explicit, biases influence judgments and behavior outside of awareness, but this evidence is rather weak and prone to alternative interpretations.

Lesson 2: Conceptual correspondence is essential for interpretations of dissociations between implicit and explicit bias.

A central issue discussed under Lesson 1 is that correlations between implicit and explicit measures can be low for various reasons that have nothing to do with lack of awareness (for a review, see Hofmann, Gschwendner, et al., 2005), including high motivation and opportunity to control the expression of activated mental contents (Fazio, 2007). Yet, even when these psychological factors are taken into account, correlations between implicit and explicit measures can be low for simple methodological reasons. In line with the correspondence principle in research on attitude-behavior relations (Ajzen & Fishbein, 1977), correlations between implicit and explicit measures tend to be higher when the two measures correspond in terms of their dimensionality and content. Yet, correlations tend to be rather low when there is little or no conceptual correspondence. For example, a meta-analysis by Hofmann, Gawronski, et al. (2005) found that implicit measures capturing relative preferences for one group over another show higher correlations to explicit measures of the same relative preferences compared to non-relative evaluations of one of the two groups. Similarly, implicit measures of racial bias using Black and White faces as stimuli tend to show higher correlations to explicit measures assessing judgments of the same faces compared to judgments of anti-discrimination policies and perceptions of racial discrimination (e.g., Payne, Burkley, & Stokes, 2008; see also Axt, in press). In general, correlations between implicit and explicit measures increase as a function of increasing correspondence between the two measures, and they decrease with decreasing correspondence (see Lesson 3 for a discussion of similar issues in research on the prediction of behavior).

Although the correspondence principle is uncontroversial among attitude researchers, its significance has been largely ignored in the literature on implicit bias. To the extent that measures of implicit and explicit bias do not correspond in terms of their target object, type of measure would be confounded with target object, rendering dissociations between the two measures ambiguous. To illustrate this problem, imagine a study in which White participants completed the Modern Racism Scale (MRS; McConahay, 1986) and an EPT using Black and White faces as primes (Fazio et al., 1995). Imagine further that the implicit measure predicted spontaneous nonverbal reactions in an interracial interaction, and the explicit measure predicted deliberate verbal behavior in the same interaction (for examples, see Dovidio et al., 2002; Fazio et al., 1995). Based on extant theories, such a finding may be interpreted as evidence for the hypothesis that implicit measures should predict spontaneous but not deliberate behavior, whereas explicit measures should predict deliberate but not spontaneous behavior (e.g., Dovidio & Gaertner, 2004; Fazio & Towles-Schwen, 1999; Strack & Deutsch, 2004; Wilson, Lindsey, & Schooler, 2000). However, in a strict sense, the finding could also be driven by the different contents of the two measures. That is, evaluations of faces might be more strongly related to spontaneous nonverbal behavior in interracial interactions regardless of whether evaluations of faces are assessed with an implicit or an explicit measure (e.g., an explicit measure asking participants to rate the faces presented in the evaluative priming task; see Payne et al., 2008). Conversely, responses to the social issues covered by the items of the MRS (e.g., perception of discrimination, evaluations of anti-discrimination policies) might be more strongly related to deliberate verbal behavior in interracial interactions regardless of whether responses to these issues are captured with the MRS or a corresponding implicit measure. Similar considerations apply to research on the incremental validity of implicit measures, which suggests that implicit measures often explain unique variance of a given outcome measure over and above explicit measures (for a review, see Perugini, Richetin, & Zogmaister, 2010). To the extent that type of measure is confounded with different target objects, such findings may speak to the incremental validity of measures assessing different contents, which may be independent of whether these measures are implicit or explicit.

The same concerns apply to studies on the determinants of implicit and explicit bias. For example, writing a counterattitudinal essay in support of

antidiscrimination policies (see Festinger & Carlsmith, 1959; Leippe & Eisenstadt, 1994) may reduce racial bias on the MRS without affecting racial bias on an IAT. However, different from the conclusion that cognitive dissonance changes explicit but not implicit bias (see Gawronski & Strack, 2004), the obtained dissociation may also be due to the different contents of the two measures. That is, writing a counterattitudinal essay in support of antidiscrimination policies may change attitudes toward antidiscrimination policies regardless of whether these attitudes are assessed with an explicit or an implicit measure. Conversely, writing a counterattitudinal essay in support of antidiscrimination policies may leave evaluations of Black and White faces unaffected regardless of whether these evaluations are assessed with an implicit or an explicit measure.

An important aspect in this context is the difference between responses to categories and responses to exemplars of a given category. A common practice in research on implicit and explicit bias is to use images of exemplars (e.g., Black and White faces as primes in an evaluative priming task) as target stimuli in the implicit measure and to assess evaluations of the relevant categories in the explicit measure (e.g., feeling thermometer or semantic differential ratings of the categories *Black people* and *White people*). Although it seems reasonable to assume that a person's responses to the exemplars of a given category are related to that person's responses to the category in general, evaluations of exemplars and categories are conceptually distinct constructs (Ledgerwood, Eastwick, & Smith, in press). Thus, studies using exemplars as target objects in implicit measures and categories as target objects in explicit measures include a confound between type of measure and target object, rendering any dissociations between the two measures ambiguous.

The non-trivial implications of this confound can be illustrated with a reanalysis of data by Gawronski, Peters, Brochu, and Strack (2008, Study 3). The study included an affect misattribution procedure (AMP; Payne et al., 2005) using Black and White faces as primes, a feeling thermometer assessing evaluations of the categories *Black people* and *White people*, and likeability ratings of the Black and White faces used as primes in the AMP. AMP scores of racial bias showed a significant positive correlation with racial bias in the likeability ratings of the faces (r = .45, p < .001), but AMP scores were unrelated to racial bias in feeling thermometer ratings of the categories (r = -.09, p = .40). Interestingly, racial bias in the likeability ratings of the faces were also unrelated to racial bias in feeling thermometer ratings of the categories (r = .07, p = .51). Together, these results suggest that, counter the idea that dissociations between AMP scores of racial bias and feeling thermometer preferences reflect genuine differences between implicit and explicit bias, such dissociations are (at least partly) rooted in the difference between responses to exemplars versus categories.

Some readers might wonder about the implications of these differences for research using the IAT, which seems to be sensitive to both the specific exemplars presented in the task and the particular categories applied to a given exemplar (e.g., Bluemke & Friese, 2006; De Houwer, 2001; Govan & Williams, 2004; Mitchell, Nosek, & Banaji, 2003). A reanalysis of data by Gawronski, Morrison, Phills, and Galdi (2017, Study 2) supports the idea that IAT scores reflect responses to both exemplars and categories. In their study, IAT scores of racial bias showed significant positive correlations with likeability ratings of the faces used in the IAT (r = .37, p < .001) and with feeling thermometer ratings of the categories (r = .38, p < .001). Moreover, the relation to either measure remained statistically significant after controlling for the respective other, in that IAT scores were still positively related to likeability ratings of the faces after controlling for feeling thermometer ratings of the categories (r = .17, p = .032) and to feeling thermometer ratings of the categories after controlling for likeability ratings of the faces (r = .20, p = .011). These findings suggest that any finding with the IAT (e.g., experimental effect on IAT scores; correlation between IAT scores and another measure) could be driven by either exemplar or category responses. This ambiguity makes it necessary to include explicit measures of both exemplar and category responses to avoid incorrect interpretations of potential dissociations in terms of features of the measure (i.e., implicit vs. explicit) rather than target objects (i.e., exemplars vs. categories).

Although the distinction between responses to categories and responses to exemplars raises important questions about the processes underlying their relation (e.g., role of inductive inferences in bottom-up effects of exemplar responses on category responses; role of deductive inferences in top-down effects of category responses on exemplar responses; see Ledgerwood et al., in press), it is just one example of how confounds between type of measure and measured contents lead to ambiguities in the interpretation of empirical findings. Another example is the difference between evaluations of objects and behaviors. Different from the emphasis on evaluations of behaviors in traditional theories of attitude-behavior relations (see Ajzen, Fishbein, Lohman, & Albarracín, 2019), most implicit measures capture evaluations of objects rather than evaluations of behaviors toward those objects. Thus, to the extent that implicit measures are designed to capture evaluations of objects (e.g., evaluations of a Muslim political candidate) and explicit measures are designed to capture evaluations of behaviors toward these objects

(e.g., evaluations of supporting a Muslim political candidate), type of measure (implicit vs. explicit) would be confounded with different contents (objects vs. behaviors), rendering dissociations between the two measures ambiguous.

Implications

Lesson 2 suggests that conceptual correspondence is essential for understanding the unique psychological properties of implicit and explicit bias. To the extent that an implicit measure has little or no conceptual correspondence with an explicit measure, their relation can be expected to be low for simple methodological reasons (Ajzen & Fishbein, 1977). In such cases, it would be premature to interpret their weak relation as evidence for the hypothesis that implicit and explicit measures capture distinct constructs (e.g., Bar-Anan & Vianello, 2018; Nosek & Smyth, 2007). Similarly, if type of measure is confounded with different contents, any finding suggesting distinct antecedents or distinct predictive relations remains ambiguous, because the obtained dissociation could be due to either (1) the implicit versus explicit nature of the measures or (2) the different contents of the two measures. Given the large proportion of studies that confounded type of measure with different contents (for a discussion, see Payne et al., 2008), a sobering conclusion is that, despite more than 20 years of research, many important questions about the properties of implicit versus explicit bias still require future research to provide unambiguous answers. At this point, it is entirely possible that several findings suggesting unique psychological properties of implicit versus explicit bias turn out to be independent of the distinction between implicit and explicit measures, and instead reflect differences in terms of the measured contents (e.g., responses to categories vs. responses to exemplars). Thus, to provide more compelling evidence for genuine differences between implicit and explicit bias, it is essential to utilize measures that correspond in terms of the measured contents (e.g., Payne et al., 2008). To the extent that previously obtained dissociations between implicit and explicit bias disappear when their respective contents are held constant, claims about functional differences between implicit and explicit bias would be empirically unfounded.

Lesson 3: There is no basis to expect strong unconditional relations between implicit bias and behavior.

A debated issue in the literature on implicit bias is whether it predicts behavior. Although numerous individual studies have found significant relations between implicit measures and behavioral outcomes (for reviews, see Friese, Hofmann, & Schmitt, 2008; Perugini et al., 2010), the average effect sizes obtained in meta-analyses tend to be rather small, with correlations ranging from .12 to .28 (Cameron et al., 2012; Greenwald, Poehlman, Uhlmann, & Banaji, 2009; Kurdi et al., in press; Oswald et al., 2013). Although some researchers suggested that statistically small relations between implicit bias and behavior could nevertheless have large societal effects (Greenwald, Banaji, & Nosek, 2015), the obtained average correlations are certainly disappointing for researchers who aim to use implicit measures to improve the prediction of behavior at the individual level.

Critics have interpreted these findings as evidence for fundamental flaws of implicit measures (e.g., Blanton & Jaccard, 2017; Mitchell, 2018). However, it is important to keep in mind that not a single theory in this area predicts strong zero-order relations between implicit measures and behavioral criteria (e.g., Dovidio & Gaertner, 2004; Fazio & Towles-Schwen, 2007; Strack & Deutsch, 2004, Wilson et al., 2000). Although these theories differ in many important regards, they agree on the broader assumption that predictive relations between attitude measures and behavior depend on the correspondence between the processing conditions of the attitude measurement and the processing conditions of the to-be-predicted behavior (for a detailed discussion, see Gawronski & De Houwer, 2014; Fazio, 2007). Thus, given that implicit measures involve highly constrained processing conditions, implicit measures should be more likely to predict behaviors performed under similar processing conditions (i.e., unintentional behavior resulting from low deliberation) compared to behaviors performed under dissimilar processing conditions (i.e., intentional behavior resulting from high deliberation). Conversely, given that the processing conditions of explicit measures do not have any such constraints, explicit measures should be more likely to predict behaviors performed under unconstrained processing conditions (i.e., intentional behavior resulting from high deliberation) compared to behaviors performed under constrained processing conditions (i.e., unintentional behavior resulting from low deliberation).

Based on this general hypothesis, a substantial number of studies investigated whether predictive relations of implicit and explicit measures to behavior depend on the type of behavior that is predicted, the conditions under which the to-be-predicted behavior is performed, and characteristics of the person who is performing the to-be-predicted behavior (for a review, see Friese et al., 2008). The three general findings of these studies are that: (1) implicit measures outperform explicit measures in the prediction of spontaneous behavior, whereas explicit measures outperform implicit measures in the prediction of deliberate behavior (e.g., Asendorpf, Banse, & Mücke, 2002; Dovidio et al., 2002; Fazio et al., 1995); (2) implicit

measures outperform explicit measures in the prediction of behavior performed under conditions that impair cognitive deliberation, whereas explicit measures outperform implicit measures in the prediction of behavior under conditions that permit cognitive deliberation (e.g., Friese, Hofmann, & Wänke, 2008; Hofmann, Gschwendner, Castelli, & Schmitt, 2008; Hofmann, Rauch, & Gawronski, 2007); and (3) implicit measures outperform explicit measures in the prediction of behavior by individuals with a disposition linked to low deliberation (e.g., low working memory capacity, intuitive thinking style), whereas explicit measures outperform implicit measures in the prediction of behavior by individuals with a disposition linked to high deliberation (e.g., high working memory capacity, deliberate thinking styles) (e.g., Hofmann, Gschwendner, Friese, Wiers, & Schmitt, 2008; Richetin, Perugini, Adjali, & Hurling, 2007).

Depending on these theoretically derived moderators, behavior should show stronger predictive relations to either implicit or explicit evaluations. Thus, to the extent that these moderators are ignored and predictive relations are averaged across different kinds of behaviors, different experimental conditions, and participants with different dispositions, the obtained average correlations should be positive but relatively small overall, as found in every published meta-analysis on the prediction of behavior with implicit measures (Cameron et al., 2012; Greenwald et al., 2009; Kurdi et al., in press; Oswald et al., 2013). Not a single meta-analysis has found a non-significant average correlation close to zero or a negative correlation. Moreover, meta-analyses that coded predictive relations obtained within a given study for theoretically derived moderators (e.g., when a given study included measures of both spontaneous and deliberate behavior) found patterns consistent with the assumptions of extant theories, in that implicit measures showed stronger relations to behavior under constrained processing conditions compared to behavior under unconstrained processing conditions (Cameron et al., 2012).

However, there is also some evidence that poses a challenge to the moderator hypotheses of extant theories. Contrary to the idea that implicit measures should show stronger relations to spontaneous compared to deliberate behavior, several meta-analysis that coded the predictive relations obtained in different studies for theoretically derived moderators found no relation between processing conditions and the size of predictive relations (e.g., Cameron et al., 2012; Kurdi et al., in press; Greenwald et al., 2009). In other words, whereas processing conditions *within* studies did show the hypothesized moderation of predictive relations, processing conditions *between* studies did not.

There are at least two potential explanations for this paradox. First, it is possible that the assumptions of extant theories are incorrect, and that the obtained moderation within studies is the product of false positives in the individual studies that included direct comparisons of processing conditions. Second, it is possible that the assumptions of extant theories are correct, and that the failure to detect a significant moderation in between-study comparisons is due to error variance resulting from procedural differences between studies. In line with the second interpretation, Cameron et al. (2012) argued that between-study comparisons aggregate across predictor and outcome measures that differ in numerous ways other than the coded variables, which can undermine the detection of actually existing effects.

One important factor in this regard is the reliability of the behavioral criterion measures. Although extant theories suggest a central role of behavior-related, situation-related, and person-related factors, previous meta-analyses have focused predominantly on behavior-related factors, such as the spontaneous versus deliberate nature of the to-be-predicted behavior (e.g., nonverbal vs. verbal behavior). To the extent that the employed measures of deliberate behavior are more reliable than the employed measures of spontaneous behavior (the latter of which are often assessed with a single item), predictive relations should be generally stronger for deliberate compared to spontaneous behavior (regardless of the predictor). In this case, implicit and explicit measures should show asymmetric relations to spontaneous versus deliberate behavior that are consistent with the hypotheses of extant theories about explicit measures, but inconsistent with their hypotheses about implicit measures. For explicit measures, the described asymmetry in the reliability of behavioral criteria should produce strong relations to deliberate behavior (because of matching processing conditions with a reliable behavioral criterion) and relatively weak or non-significant relations to spontaneous behavior (because of mismatching processing conditions with an unreliable behavioral criterion). In contrast, for implicit measures, the described asymmetry in the reliability of the behavioral criteria should produce relatively weak relations to both spontaneous behavior (because of low reliability of the behavioral measure) and deliberate behavior (because of mismatching processing conditions). Indeed, this asymmetric pattern of predictive relations emerged in every meta-analysis that compared predictive relations of implicit and explicit measures to spontaneous versus deliberate behavior on a between-study basis (Cameron et al., 2012; Greenwald et al., 2009; Kurdi et al., in press). Although some authors interpreted this pattern as evidence against the hypotheses of extant theories (e.g., Kurdi et al., in press; Greenwald et al., 2009), it

would be consistent with these theories to the extent that the measures of spontaneous behavior were less reliable than the measures of deliberate behavior (e.g., when spontaneous behavior was measured with a single item and measures of deliberate behavior included multiple items).

Another important issue in the evaluation of the weak predictive relations obtained in meta-analyses is that strong relations should be limited to cases in which implicit measures have high conceptual correspondence with the behavioral criterion (see Lesson 2). To the extent that conceptual correspondence between the two measures is low, their relation should be weak regardless of the moderators proposed by extant theories (see Ajzen & Fishbein, 1977). For example, in a study by Amodio and Devine (2006), a measure of implicit evaluative bias was significantly related to participants' desire to befriend a racial outgroup member, but not to their expectations about the outgroup member's performance on a trivia task (but see Supplemental Materials of Oswald et al., 2013, for a potential error in the relations reported for implicit evaluative bias). Conversely, a measure of implicit stereotypical bias was significantly related to participants' expectations about the outgroup member's performance on a trivia task, but not to their desire to befriend the outgroup member. In line with these findings, a recent meta-analysis by Kurdi et al. (in press) found relatively large relations between IAT measures and intergroup behavior when the two measures had high conceptual correspondence (average correlation of r = .37). However, IAT measures showed no significant relation to intergroup behavior when conceptual correspondence was low (average correlation of r = .02).

Together, these considerations suggest that average relations obtained in meta-analyses ignore important complexities in the prediction of behavior with implicit and explicit measures. Strong predictive relations can be expected to emerge only when (1) there is high conceptual correspondence between the predictor measure and the behavioral criterion, and (2) the processing conditions of the predictor measure match to the processing conditions of the to-be-predicted behavior. Thus, when predictive relations are averaged in a single meta-analytic effect size, implicit measures should show significant positive, but relatively weak, relations to behavior, as found in every meta-analysis on the prediction of behavior with implicit measures (Cameron et al., 2012; Greenwald et al., 2009; Kurdi et al., in press; Oswald et al., 2013). Of course, there is no guarantee that the hypotheses of extant theories are correct, and that future studies and meta-analytic reviews will support the predictions derived from these theories. However, a focus on unconditional zero-order relations in the prediction of behavior can be criticized for ignoring the current state of theory and research on attitude-behavior relations. On the one hand, attempts to show large unconditional relations between implicit measures and behavior seem unlikely to succeed, given that there is no theoretical and methodological basis to expect large unconditional relations. On the other hand, criticism of implicit measures for showing relatively weak average relations to behavior seems premature, given that predictive relations can be expected to be relatively weak when theoretical and methodological moderators are ignored.

Implications

Lesson 3 suggests that there is no basis to expect strong unconditional relations between implicit bias and behavior. Thus, research on the prediction of behavior would benefit from focusing on moderators of predictive relations rather than zero-order correlations between implicit bias and behavior. Although extant theories differ in many important regards, they agree on the general assumption that predictive relations between attitudes and behavior should depend on the correspondence between the processing conditions of the attitude measurement and the processing conditions of the to-be-predicted behavior (e.g., Dovidio & Gaertner, 2004; Fazio & Towles-Schwen, 1999; Strack & Deutsch, 2004; Wilson et al., 2000). Based on this assumption, predictive relations of implicit and explicit measures to behavior should depend on the type of behavior that is predicted, the conditions under which the to-be-predicted behavior is performed, and characteristics of the person who is performing the to-be-predicted behavior. Although the findings of several individual studies support these assumptions (for a review, see Friese et al., 2008), future research may be more successful in convincing skeptics by following recently established best practices to avoid false positives (e.g., sufficiently large sample sizes, preregistration, independent replication, etc.). Because differences in the reliability of measurement instruments can distort the patterns of dissociations obtained with implicit and explicit measures, an important issue in this endeavor is to ensure comparable reliabilities of the employed predictor measures as well as the measures of the to-be-predicted outcomes. Finally, because low conceptual correspondence should lead to low predictive relations regardless of the moderators proposed by extant theories (see Lesson, 2), the contents of the predictor measures should correspond to the contents of the to-be-predicted behaviors. Of course, there is no guarantee that such studies will support the predictions derived from extant theories. However, research focusing exclusively on unqualified zero-order correlations could be criticized for making a rather small scientific contribution, because it ignores the current state of the field.

Lesson 4: Implicit bias is less (not more) stable over time than explicit bias.

Although Lesson 3 suggests that implicit measures might be valuable tools for the prediction of behavior if the identified moderators are taken into account, there is a more fundamental issue that can undermine the utility of implicit measures in predicting future behavior. Counter to the widespread assumption that the constructs captured by implicit measures are highly stable, findings of several longitudinal studies suggest that implicit measures tend to show lower test-retest correlations compared to explicit measures, even when the two kinds of measures show comparable estimates of internal consistency. For example, across two longitudinal studies that compared the temporal stability of implicit and explicit measures over a period of one to two months in three content domains (i.e., racial attitudes, political attitudes, self-concept), Gawronski et al. (2017) found a weighted average stability of $r = .54$ for implicit measures and a weighted average stability of $r = .75$ for explicit measures (for similar findings, see Bosson, Swann, & Pennebaker, 2000; Cunningham, Preacher & Banaji, 2001; Galdi, Arcuri, & Gawronski, 2008; Galdi, Gawronski, Arcuri, & Friese, 2012; Rae & Olson, 2018). These results suggest that a person's score on an implicit measure today provides limited information about this person's score on the same measure at a later time. Needless to say, such temporal fluctuations can be detrimental if the goal is to predict future behavior based on the scores of an implicit measure obtained at an earlier time. Explicit measures fare better in this regard, in that they show significantly higher stability over time compared to implicit measures. From this perspective, explicit measures can be expected to be superior predictors of future behavior regardless of the moderators hypothesized by extant theories (see Lesson 3), simply because explicit measures tend to show less temporal fluctuations than implicit measures.

Although the low temporal stability of implicit measures can undermine their usefulness in predicting future behavior, this limitation does not necessarily question their construct validity, as suggested by some critics of implicit measures (e.g., Mitchell, 2018). From a psychometric view, low temporal stability simply suggests a low proportion of stable trait variance. Yet, in contrast to widespread interpretations of implicit measures as pure indicators of temporally stable traits, a considerable proportion of temporally fluctuating variance may reflect momentary states. The latter conclusion is consistent with studies that used latent state-trait analysis to decompose the contributions of situation-related and person-related factors in implicit measures (e.g., Dentale, Veccione, Ghezzi, & Barbaranelli, in press; Koch, Ortner, Eid, Caspers, & Schmitt, 2014; Lemmer, Gollwitzer, & Banse, 2015; Schmukle & Egloff, 2005). Consistent with the findings of these studies, some theories suggest that implicit measures reflect the momentary activation of associations in memory, which depends on situational factors over and above a person's chronic structure of associations in memory (e.g., Gawronski & Bodenhausen, 2006, 2011). Thus, although temporal fluctuations in the momentary activation of associations can be detrimental for the prediction of future behavior via implicit measures, this limitation does not necessarily question the construct validity of implicit measures as indicators of a person's thoughts at the time of measurement. Indeed, it would seem premature to dismiss a measure that is supposed to capture what is on a person's mind in a given moment simply because the measure shows different results over time. After all, a person's thoughts in a given moment are determined not only by personal, but also by situational factors.

Nevertheless, the fact that implicit measures show relatively low stability over time conflicts with a common narrative in the literature, according to which (1) a person's score on an implicit measure reflects a trait-like characteristic of that person and (2) these traits are acquired early in childhood and remain stable over the course of development (e.g., Baron & Banaji, 2006; Rudman, Phelan, & Heppen, 2007). Although the obtained test-retest correlations are consistent with the idea that implicit measures are at least partly influenced by trait-like characteristics, the overall size of these correlations suggest that situation-related factors have a considerable impact on implicit measures over and above trait-related factors. Moreover, given that a person's scores on the same implicit measure fluctuate considerable over a few weeks (e.g., Bosson et al., 2000; Cunningham et al., 2001; Gawronski et al., 2017; Galdi et al., 2008; Rae & Olson, 2018), claims that these scores reflect trait-like characteristics acquired during childhood seem difficult to reconcile with the available evidence (see also Castelli, Carraro, Gawronski, & Gava, 2010).

The low temporal stability of implicit measures also raises the question of why children as young as 6 years show levels of implicit biases that are indistinguishable from the ones shown by adults (e.g., Banse, Gawronski, Rebetez, Gutt, & Morton, 2010; Baron & Banaji, 2006). Payne, Vuletich, and Lundberg (2017) argued that this paradox could be resolved by assuming that (1) implicit biases reflect currently accessible concepts and (2) concept accessibility is primarily determined by environmental factors (see also Dasgupta, 2013). Thus, to the extent that adults and children are exposed to the same environmental factors, they should show similar average levels of implicit bias, as found in several developmental studies (e.g., Banse et al., 2010; Baron & Banaji, 2006; but see Degner & Wentura, 2010). This explanation reconciles the low

temporal stability of implicit measures with the finding that children and adults show similar average levels of implicit bias. Low temporal stability at the individual level is explained by the strong impact of transient situational factors at the individual level, and comparable average levels of implicit bias among children and adults are explained by the fact that children and adults tend to live in the same cultural environments. However, the strong emphasis on situational factors in this explanation implies the possibility that even the temporally stable component of implicit biases is the product of situational factors (see Payne et al., 2017). To the extent that people's cultural environments are at least somewhat stable and consistent over time, the obtained level of stable variance in implicit measures may reflect the relative stability of people's environments rather than trait-like characteristics of individuals (Lord & Lepper, 1999; Schwarz, 2007). Although radical situationist interpretations of implicit bias seem difficult to reconcile with evidence for mutual interactions between person-related and situation-related factors (see Lesson 5), the possibility that temporally stable variance may reflect stable environments poses an even greater challenge to the idea that implicit bias scores provide diagnostic information about traits (see also Livingston, 2002).[5]

Implications

A common narrative in research on implicit bias suggests that (1) a person's score on an implicit measure reflects a trait-like characteristic of that person and (2) these traits are acquired early in childhood and remain stable over the course of development. These assumptions are difficult to reconcile with a substantial body of evidence showing that implicit biases tend to fluctuate considerably over time, and in fact are less stable over time compared to explicit biases. Although these findings do not necessarily question the construct validity of implicit measures, they suggest an interpretation of implicit biases that is fundamentally different from the mainstream narrative. Different from dominant interpretations of implicit biases as reflecting temporally stable characteristics of a person, the available evidence suggests that implicit measures capture both traits and states. This conclusion is relevant not only for conceptual interpretations of implicit biases; it also has important implications for research on the prediction of behavior and the antecedents of implicit biases. On the one hand, the low temporal stability of implicit biases pose a major challenge for the prediction of behavior over time. On the other hand, the contribution of transient states suggests that intervention-related changes in implicit bias may reflect short-lived changes in the state of a given individual rather than temporally stable changes in that person's traits.

Lesson 5: Context matters fundamentally for the outcomes obtained with implicit bias measures.

The conclusions of Lesson 4 imply that contextual factors are essential for understanding the outcomes obtained with implicit measures. In fact, the available evidence suggests that contextual factors determine virtually every finding with implicit measures, including (1) their overall scores, (2) their temporal stability, (3) the prediction of future behavior, and (4) the effectiveness of interventions. Although the significance of contextual factors has been identified in the early years of research with implicit measures (Blair, 2002), contextual thinking has still not penetrated the mainstream narrative about implicit bias.

With regard to the overall scores obtained with implicit measures, a substantial body of research demonstrated that implicit measures are highly sensitive to a broad range of contextual factors (for a review, see Gawronski & Sritharan, 2010). Examples of contextual factors that have been shown to influence implicit bias include recently encountered exemplars of a given category (e.g., Dasgupta & Asgari, 2004; Dasgupta & Greenwald, 2001), the environment in which a given target person is encountered (e.g., Maddux, Barden, Brewer, & Petty, 2005; Wittenbrink, Judd, & Park, 2001), contextually salient categories (e.g., Kühnen, Schiessl, Bauer, Paulig, Pöhlmann, & Schmidthals, 2001; Mitchell et al., 2003), the social role of the perceiver (e.g., Richeson & Ambady, 2001, 2003), and incidental emotional states of the perceiver (e.g., Dasgupta, DeSteno, Williams, & Hunsinger, 2009; DeSteno, Dasgupta, Bartlett, & Cajdric, 2004). Based on a review of these findings, Gawronski and Bodenhausen (2006) argued that exposure to a given stimulus does not activate all components of the stored representation of that stimulus. Instead, activation is limited to a subset of stored information, and contextual cues influence which aspects of the representation are activated in response to given stimulus (see also Ma, Correll, & Wittenbrink, 2006).

With regard to context effects on the temporal stability of implicit bias, there is evidence that implicit measures show greater test-retest correlations to the extent that (1) meaningful context cues constrain the

[5] Similar to historical debates in personality psychology (Snyder & Ickes, 1985), a potential qualification to this conclusion is that people may at least partly chose their environments. In this case, implicit measures could provide indirect information about trait-related characteristics even if they exclusively reflect situational influences (for evidence regarding mutually reinforcing effects of person-related and situation-related factors, see Galdi et al., 2012).

activation of stored information and (2) these context cues are consistent over time. In a largely neglected study on this issue, Gschwendner, Hofmann, and Schmitt (2008) found rather low levels of stability in implicit bias over a period of two weeks when they used a standard variant of the IAT (r = .29). However, temporal stability of implicit bias over the same period was significantly higher when the measure included background images to provide meaningful information about the context of the target stimuli (r = .72).[6] These findings suggest that a person's level of implicit bias fluctuates over time in the absence of strong contextual constraints. However, implicit bias seems to be quite stable over time to the extent that contextual constraints are strong and consistent across measurements.

In addition to demonstrating the impact of contextual factors on the temporal stability of implicit measures, Gschwendner et al.'s (2008) findings also have important implications for the prediction of future behavior with implicit measures. Because implicit measures tend to show considerable fluctuation over time in the absence of strong contextual constraints (e.g., Bosson et al., 2000; Cunningham et al., 2001; Gawronski et al., 2017, Galdi et al., 2008; Rae & Olson, 2018), it seems unrealistic to expect to strong relations between previously administered implicit measures and future behavior under such conditions. After all, it seems unlikely that a measure would predict future behavior if the scores on the measure today are weakly related to the scores on the same measure at a later time (see Lesson 4). Yet, predictive relations to future behavior may be higher to the extent that scores on the predictor measure are stable over time (for a discussion, see Ajzen & Fishbein, 1980). Thus, given that implicit measures show considerable levels of temporal stability when contextual constraints are strong and consistent across measurements, the latter conditions may also increase their predictive relations to future behavior.

A final issue concerns the role of contextual factors in understanding the effectiveness of interventions to change implicit bias. A central question in the literature on bias intervention is whether the effects of a given intervention remain stable over time. In a large-scale study that compared the effectiveness of 17 interventions to reduce implicit bias, Lai et al. (2014) found considerable differences in the immediate effects of the tested interventions, in that some interventions effectively reduced implicit bias, whereas others did not. However, a follow-up study comparing the 9 most effective interventions revealed that not a single one of them produced stable reductions over time (Lai et al., 2016). Although every intervention reduced implicit bias immediately after the intervention, implicit bias went back to pre-intervention baselines for all 9 interventions.

One potential interpretation of this finding is that the tested interventions merely influenced the subset of stored information that was activated in response to a given stimulus, similar to the reviewed effects of contextual factors (see Gawronski & Sritharan, 2010). In this case, the obtained effects on implicit bias would reflect fleeting changes in the momentary activation of stored information rather than changes in the stored representation itself (see Lesson 4). Yet, an alternative interpretation is that the tested interventions effectively changed the stored representation, but these changes were limited to the context in which the intervention occurred. Research inspired by the notion of contextual renewal in animal learning (see Bouton, 2004) suggests that the effects of counterattitudinal information are sometimes limited to the context in which the counterattitudinal information was learned (for a review, see Gawronski, Rydell, De Houwer, Brannon, Ye, Vervliet, & Hu, 2018). The typical pattern obtained in this research is that counterattitudinal information determines evaluative responses in the context in which the counterattitudinal information was learned, whereas initial attitudinal information continues to influence responses in any other context, including the context in which the initial attitudinal information was learned or novel contexts in which the target object has not been encountered before (e.g., Brannon & Gawronski, 2018; Gawronski, Rydell, Vervliet, & De Houwer, 2010; Gawronski, Ye, Rydell, & De Houwer, 2014; Rydell & Gawronski, 2009; Ye, Tong, Chiu, & Gawronski, 2017).

Because Lai et al.'s (2016) participants completed the study online and there was no control over the context in which participants completed the two sessions, it is possible that participants completed the delayed follow-up measurement in a context that was different from the context of the intervention and the immediate assessment of implicit bias. In this case, the reduced effectiveness of the 9 interventions in influencing implicit bias at the follow-up measurement may have been due to a change in context rather than low stability of changes over time. That is, a given intervention may be effective in producing long-term changes in implicit bias within the context in which the intervention occurred, but the effects of the intervention may be limited in the sense that they do not generalize across contexts. Conversely, even if a given intervention effectively reduces implicit bias within the same context over time, the effectiveness of the

[6] Similar findings were obtained for an Implicit Association Test designed to measure the implicit self-concept of anxiety.

intervention could be limited in the sense that the observed reduction is limited to the context in which the intervention occurred. Thus, to establish the effectiveness of a given intervention, it is important to include not only delayed follow-up measurements, but also measurements in contexts that are different from the one in which the intervention took place (Gawronski & Cesario, 2013).

At a broader level, a central implication of the reviewed findings is that implicit biases might be better understood in terms of complex person-by-situation interactions rather than exclusive effects of person-related or situation-related factors (Mischel & Shoda, 1995). A person may show different responses to the same stimulus depending on the context in which the stimulus is encountered. Conversely, different people may show different responses to a given stimulus within same context, and these context-specific individual differences may be relatively stable over time. Theoretically, these patterns can be explained as the interactive products of (1) the pre-existing structure of associations in memory (person-related factor) and (2) the overall configuration of input stimuli (situation-related factor). The two factors constrain each other in the sense that (1) the pre-existing structure of associations in memory constrains the contents that are activated in response to a given stimulus and (2) context stimuli constrain which pre-existing associations are activated in response to a target stimulus (Gawronski & Bodenhausen, 2017).

Implications

Lesson 5 suggests that context matters fundamentally for the outcomes obtained with implicit measures, including (1) their overall scores, (2) their temporal stability, (3) the prediction of future behavior, and (4) the effectiveness of interventions. Related to the notion that implicit biases reflect both traits and states (see Lesson 4), contextual factors have been found to influence overall levels of implicit bias. Moreover, strong contextual constraints have been found to increase the temporal stability of implicit biases, suggesting a major role for person-by-situation interactions. Further, the higher stability of implicit biases under conditions of strong contextual constraints suggests that strong relations between implicit bias and future behavior require consistent contextual constraints over time. Finally, the notion of contextual renewal suggests that, even if intervention-related changes are temporally stable within the context in which the intervention occurred, the observed changes may not generalize to other contexts. Future research on implicit bias would benefit from paying more attention to these multiple ways by which contextual factors can influence the outcomes obtained with implicit measures.

Lesson 6: Implicit measures do not provide process-pure reflections of bias.

A final lesson is that implicit measures do not provide process-pure reflections of a focal construct (e.g., racial bias). Like any psychological measure, variance in the scores obtained with implicit measures (X) comprise variance reflecting the construct of interest (C), systematic error (SE), and random error (RE), which can be depicted in the equation:

$$X = C + SE + RE$$

Somewhat surprisingly, this widely accepted insight is rarely considered in research on implicit bias, which can lead to inaccurate conclusions about its psychological properties.

One important issue in this regard is that implicit measures based on response interference are strongly influenced by executive control processes over and above the impact of dominant response tendencies reflecting bias (Conrey, Sherman, Gawronski, Hugenberg, & Groom, 2005). For example, in an IAT designed to measure racial bias, negativity toward African Americans may elicit a prepotent tendency to press the "negative" key in response to Black faces. This tendency should facilitate quick and accurate responses when the response key for negative stimuli is the same as the one for Black faces. In contrast, quick and accurate responses should be inhibited when the response key for negative stimuli is different from the one for Black faces. Importantly, the speed and accuracy of responses in the latter block is not only influenced by the strength of the prepotent tendency to press the "negative" key (presumably reflecting the degree of negativity toward African Americans). Speed and accuracy in this block also depend on executive control processes, given that participants have to suppress their prepotent response tendency in order to provide the correct response. Because executive control varies across individuals and contextual factors, variance in IAT scores not only comprises variance in the construct of interest (e.g., racial bias), but also variance reflecting systematic error (i.e., executive control).

This insight has important implications for both experimental and correlational research using implicit measures. For example, to the extent that an experimental manipulation influences measurement scores on an IAT designed to measure racial bias, the obtained effect may reflect either (1) a difference in racial bias or (2) a difference in executive control, or both (see Sherman et al., 2008). Moreover, to the extent that given manipulation influences racial bias and executive control in ways that compensate each other (e.g., higher levels of racial bias compensated by higher levels of executive control), the experimental manipulation may show a null effect on traditional IAT scores (see Sherman et al., 2008). Similar concerns

apply to research using correlational designs. For example, if measurement scores on an IAT designed to measure racial bias show a significant correlation with a criterion measure (e.g., behavior), this correlation could be driven by either (1) shared variance in the construct of interest (e.g., racial bias) or (2) shared variance in systematic error (e.g., executive control), or both.

One potential way to resolve these ambiguities is the use of formal modeling procedures to analyze the data obtained with an implicit measure (for a review, see Sherman, Klauer, & Allen, 2010). One example is Conrey et al.'s (2005) quad-model, which allows researchers to quantify the contributions of four qualitatively distinct processes to IAT performance: activation of an association (*AC*), detection of the correct response required by the task (*D*), success at overcoming associative bias (*OB*), and guessing (*G*). An alternative strategy is to replicate a given finding with implicit measures that have distinct sources of systematic error, as can be expected for implicit measures that are based on different underlying processes (see Gawronski, Deutsch, LeBel, & Peters, 2008). For example, in contrast to the response interference mechanism underlying the IAT and evaluative priming (De Houwer, 2003b), the AMP is based on a misattribution mechanism that involves sources of systematic error that are distinct from the ones affecting scores on the IAT and evaluative priming (Gawronski & Ye, 2014). Thus, successful replications with two types of implicit measures provide a stronger basis for conclusions that a given effect is driven by the construct of interest rather than sources of systematic error (e.g., Peters & Gawronski, 2011; Prestwich, Perugini, Hurling, & Richetin, 2010).

The significance of task-specific mechanisms can be illustrated with findings, showing that the same experimental manipulation can have distinct effects on implicit measures with different underlying mechanisms (e.g., Deutsch & Gawronski, 2009; Gawronski & Bodenhausen, 2005; Gawronski, Cunningham, LeBel, & Deutsch, 2010). For example, in a series of studies by Gawronski et al. (2010), participants completed an EPT using Black and White faces of either young or old age as primes. Half of the participants were instructed to count the number of Black and White faces presented in the task; the remaining half were asked to count the number of young and old faces (see Olson & Fazio, 2003). Gawronski et al. found reliable priming effects of implicit race bias when participants paid attention to race, but not when they paid attention age. Conversely, reliable priming effects of implicit age bias emerged only when participants paid attention to age, but not when they paid attention to race. This pattern was reflected in the overall size of priming effects, their internal consistency, and their relation to corresponding measures of explicit bias. Based on extant theories (e.g., Fazio, 2007; Gawronski & Bodenhausen, 2006), this finding may be interpreted as evidence for the hypothesis that evaluative responses to a given stimulus depend on how perceivers categorize that stimulus (e.g., categorization of a young Black man in terms of race versus age). However, counter to this interpretation, the same manipulation had no significant effects on priming effects in the AMP. That is, participants who completed the AMP showed reliable priming effects of implicit race bias regardless of whether they paid attention to race or age. Similarly, participants who completed the AMP showed reliable priming effects of implicit age bias regardless of whether they paid attention to age or race. Based on earlier comparisons of priming effects in the EPT and the AMP (Deutsch & Gawronski, 2009), Gawronski et al. (2010) argued that the obtained effects on the EPT reflect attentional influences on the response interference mechanism underlying the EPT rather than genuine effects on implicit bias. Specifically, the authors argued that the response interference mechanism underlying the EPT presupposes attention to the relevant features of the primes, which is not the case for the misattribution mechanism underlying the AMP. Thus, in studies that exclusively rely on implicit measures based on response interference (for a review, see Gawronski, Deutsch, & Banse, 2011), manipulations that influence participants' attention to different features of a stimulus can lead to the incorrect conclusion that these manipulations influenced implicit bias, although the obtained differences may simply reflect effects on the response interference mechanism underlying the task.

The broader significance of these issues can be illustrated with a widely cited finding of an unpublished meta-analysis of change in implicit bias. Forscher et al. (2016) found that most procedures designed to change implicit bias were effective, although average effect sizes were rather small for many of the tested interventions. Moreover, most procedures had larger effects on implicit bias compared to behavioral measures, and there was no evidence that change in implicit bias mediated change in behavior. Based on these findings, the authors concluded that changes in implicit bias do not lead to changes in behavior, which poses a challenge to the idea that implicit bias causes discriminatory behavior (Mitchell, 2018). If implicit bias was a cause of discriminatory behavior, experimentally induced changes in implicit bias should lead to corresponding changes in discriminatory behavior, which was not the case in Forscher et al.'s (2016) meta-analysis.

Although Forscher at al.'s (2016) unpublished meta-analytic findings have become a central argument

in the criticism of research on implicit bias, the criticism is based on a number of background assumptions that seem questionable in light of the issues reviewed in the current article. First, change in implicit bias should lead to corresponding change in behavior only under specific conditions (see Lesson 3). Because Forscher et al.'s (2016) meta-analysis did not code for these conditions, it is possible that discrepant effects on implicit bias and behavior are at least partly due to a mismatch of processing conditions or lack of conceptual correspondence between measures. Second, the methodological dictum that scores obtained with implicit measures (like any other psychological measure) reflects systematic construct variance as well as systematic error variance implies the possibility that some procedures may influence measurement scores via effects on sources of systematic error (e.g., executive control) rather than the constructs of interest (e.g., racial bias). For example, procedures that tax participants' cognitive resources were found to be among the most effective procedures to influence implicit bias. However, such procedures seem more likely to influence measurement scores via reduced executive control rather than genuine changes in bias. In this case, it seems rather unlikely that the obtained effect on measurement scores would be associated with corresponding effects on a behavioral criterion measure (unless resources are also taxed for the behavioral measure).

Implications

Lesson 6 suggests that research on implicit bias would benefit from explicitly considering the methodological dictum that variance in the scores obtained with implicit measures (like any other measure) reflects (1) systematic construct variance, (2) systematic measurement error, and (3) random error. This truism implies that any effect obtained with implicit measures may be driven by the construct of interest or by measurement-related processes that are independent of the to-be-measured construct. Thus, treatments of implicit measurement scores as process-pure reflections of the to-be-measured construct can lead to incorrect conclusions about the psychological properties of implicit bias. Future research on implicit bias would benefit from directly addressing these ambiguities by (1) analyzing data with formal modeling procedures that disentangle the contributions of multiple distinct processes to measurement outcomes or (2) comparing findings across implicit measures that are based on different underlying mechanisms (or both).

Conclusion

Table 1 provides an overview of the normative implications of the six lessons reviewed in this article. Although the current analysis focused primarily on implicit bias, it is worth noting that the key points are relevant for all research using implicit measures. Moreover, many of the key points apply not only to implicit but also to explicit bias. The dominant focus on implicit bias was inspired by (1) the increasing skepticism about the value of the construct in understanding social discrimination and (2) the rather low appreciation of the six lessons in research on implicit bias compared to other areas. Together, the six lessons suggest that research on implicit bias would benefit from considering the broader literature on implicit measures as well as historical debates in research on attitudes. At the same time, dismissing the implicit bias construct as entirely irrelevant for the psychological understanding of social discrimination seems premature in light of the six lessons. Of course, previous research on implicit bias can be criticized for providing ambiguous evidence that does not permit strong conclusions of either kind. However, by following the normative implications of the six lessons, future research may directly address these ambiguities, and thereby provide a more nuanced understanding of implicit bias, its psychological characteristics, and its contribution to social discrimination. Whether this research will ultimately confirm a unique role of implicit bias over and above explicit bias is an open question, and there is no guarantee that the obtained findings will suggest an affirmative answer. However, to provide a strong basis for empirically convincing conclusions of either kind, it is essential to directly address the limitations of previous research. The normative implications of the six lessons may provide a helpful framework in this endeavor, providing the foundation for a cogent science of implicit bias.

References

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

Ajzen, I., & Fishbein, M. (1980). *Understanding attitudes and predicting social behavior*. Englewood Cliffs, NJ: Prentice-Hall

Ajzen, I., Fishbein, M., Lohmann, S., & Albarracín, D. (2019). The influence of attitudes on behavior. In D. Albarracín & B. T. Johnson (Eds.), *The handbook of attitudes* (2nd ed., pp. 197-256). New York: Routledge

Albarracín, D. & Johnson, B. T. (Eds.). (2019). *The handbook of attitudes* (2nd ed.). New York: Routledge.

Amodio, D. M., & Devine, P. G. (2006). Stereotyping and evaluation in implicit race bias: Evidence for independent constructs and unique effects on behavior. *Journal of Personality and Social Psychology, 91,* 652-661.

Arkes, H. R., & Tetlock, P. E. (2004). Attributions of implicit prejudice, or "Would Jesse Jackson 'fail' the Implicit Association Test? *Psychological Inquiry, 15*, 257-278.

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

Axt, J. R. (in press). The best way to measure explicit racial attitudes is to ask about them. *Social Psychological and Personality Science*.

Baker, A. (2018, July 15). Confronting implicit bias in the New York Police Department. *New York Times*. Retrieved from https://www.nytimes.com.

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.

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

Bar-Anan, Y., & Vianello, M. (2018). A multi-method multi-trait test of the dual-attitude perspective. *Journal of Experimental Psychology: General, 147*, 1264-1272.

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.

Baron, A. S., & Banaji, M. R. (2006). The development of implicit attitudes: Evidence of race evaluations from ages 6 and 10 and adulthood. *Psychological Science, 17*, 53-58.

Bartlett, T. (2017, January 5). Can we really measure implicit bias? Maybe not. *The Chronicle of Higher Education.* Retrieved from http://www.chronicle.com.

Blair, I. V. (2002). The malleability of automatic stereotypes and prejudice. *Personality and Social Psychology Review, 6*, 242-261.

Blanton, H., & Jaccard, J. (2017). You can't assess the forest if you can't assess the trees: Psychometric challenges to measuring implicit bias in crowds. *Psychological Inquiry, 28*, 249-257.

Bluemke, M., & Friese, M. (2006). Do irrelevant features of stimuli influence IAT effects? *Journal of Experimental Social Psychology, 42*, 163-176.

Bosson, J. K., Swann, W. B., & Pennebaker, J. W. (2000). Stalking the perfect measure of implicit self-esteem: The blind men and the elephant revisited? *Journal of Personality and Social Psychology, 79*, 631-643.

Bouton, M. E. (2004). Context and behavioral processes in extinction. *Learning and Memory, 11*, 485-494.

Brannon, S. M., & Gawronski, B. (2018). Does contextualized attitude change depend on individual differences in responses to belief-incongruent information? *Journal of Experimental Social Psychology, 78*, 148-161.

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, 16*, 330-350.

Castelli, L., Carraro, L., Gawronski, B., & Gava, K. (2010). On the determinants of implicit evaluations: When the present weighs more than the past. *Journal of Experimental Social Psychology, 46*, 186-191.

Clinton, H. (2016, September 28). *For African Americans, implicit bias can be deadly. It's time to address it*. Retrieved from http://www.Hillaryclinton.com.

Conrey, F. R., Sherman, J. W., Gawronski, B., Hugenberg, K., & Groom, C. (2005). Separating multiple processes in implicit social cognition: The quad-model of implicit task performance. *Journal of Personality and Social Psychology, 89*, 469-487.

Cunningham, W.A., Nezlek, J.B., & Banaji, M. R. (2004). Implicit and explicit ethnocentrism: Revisiting the ideologies of prejudice. *Personality and Social Psychology Bulletin, 30*, 1332-1346.

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

Dasgupta, N. (2013). Implicit attitudes and beliefs adapt to situations: A decade of research on the malleability of implicit prejudice, stereotypes, and the self-concept. *Advances in Experimental Social Psychology, 47*, 233-279.

Dasgupta, N., & Asgari, S. (2004). Seeing is believing: Exposure to counterstereotypic women leaders and its effect on the malleability of automatic gender stereotyping. *Journal of Experimental Social Psychology, 40*, 642-658.

Dasgupta, N., DeSteno, D., Williams, L. A., & Hunsinger, M. (2009). Fanning the flames of prejudice: The influence of specific incidental emotions on implicit prejudice. *Emotion, 9*, 585-591.

Dasgupta, N., & Greenwald, A. G. (2001). On the malleability of automatic attitudes: Combating automatic prejudice with images of admired and disliked individuals. *Journal of Personality and Social Psychology, 81*, 800-814.

De Houwer, J. (2001). A structural and process analysis of the Implicit Association Test. *Journal of Experimental Social Psychology, 37*, 443-451.

De Houwer, J. (2003b). A structural analysis of indirect measures of attitudes. In J. Musch & K. C. Klauer (Eds.), *The psychology of evaluation: Affective processes in cognition and emotion* (pp. 219-244). Mahwah, NJ: Erlbaum.

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

Degner, J., & Wentura, D. (2008). The extrinsic affective Simon task as an instrument for indirect assessment of prejudice. *European Journal of Social Psychology, 38*, 1033-1043.

Degner, J., & Wentura, D. (2010). Automatic prejudice in childhood and early adolescents. *Journal of Personality and Social Psychology, 98*, 356-374.

Dentale, F., Veccione, M., Ghezzi, V., & Barbaranelli, C. (in press). Applying the latent state-trait analysis to decompose state, trait, and error components of the self-esteem Implicit Association Test. *European Journal of Psychological Assessment*.

DeSteno, D. A., Dasgupta, N., Bartlett, M. Y., & Cajdric, A. (2004). Prejudice from thin air: The effect of emotion on automatic intergroup attitudes. *Psychological Science, 15*, 319-324.

Deutsch, R., & Gawronski, B. (2009). When the method makes a difference: Antagonistic effects on "automatic evaluations" as a function of task characteristics of the measure. *Journal of Experimental Social Psychology, 45*, 101-114.

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

Dovidio, J. F., Kawakami, K., & Gaertner, S. L. (2002). Implicit and explicit prejudice and interracial interaction. *Journal of Personality and Social Psychology, 82*, 62-68.

Dunton, B. C., & Fazio, R. H. (1997). An individual difference measure of motivation to control prejudiced reactions. *Personality and Social Psychology Bulletin, 23*, 316-326.

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

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

Fazio, R. H., & Towles-Schwen, T. (1999). The MODE model of attitude-behavior processes. In S. Chaiken Y. Trope (Eds.), *Dual process theories in social psychology* (pp. 97-116). New York: Guilford.

Festinger, L., & Carlsmith, J. M. (1959). Cognitive consequences of forced compliance. *Journal of Abnormal and Social Psychology, 58*, 203-210.

Fiedler, K., Messner, C., & Bluemke, M. (2006). Unresolved problems with "I", the "A", and the "T": A logical and psychometric critique of the Implicit Association Test (IAT). *European Review of Social Psychology, 17*, 74-147.

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*, 285-338.

Friese, M., Hofmann, W., & Wänke, M. (2008). When impulses take over: Moderated predictive validity of explicit and implicit attitude measures in predicting food choice and consumption behavior. *British Journal of Social Psychology, 47*, 397-419.

Friese, M., Loschelder, D. D., Gieseler, K., Frankenbach, J., & Inzlicht, M. (in press). Is ego depletion real? An analysis of arguments. *Personality and Social Psychology Review*.

Forscher, P. S., Lai, C. K., Axt, J. R., Ebersole, C. R., Herman, M., Devine, P. G., & Nosek, B. A. (2016). *A meta-analysis of change in implicit bias*. Unpublished manuscript.

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.

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

Galdi, S., Gawronski, B., Arcuri, L., & Friese, M. (2012). Selective exposure in decided and undecided individuals: Differential relations to automatic associations and conscious beliefs. *Personality and Social Psychology Bulletin, 38*, 559-569.

Gawronski, B., & Bodenhausen, G. V. (2005). Accessibility effects on implicit social cognition: The role of knowledge activation and retrieval experiences. *Journal of Personality and Social Psychology, 89*, 672-685.

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

Gawronski, B., & Bodenhausen, G. V. (2011). The associative-propositional evaluation model:

Theory, evidence, and open questions. *Advances in Experimental Social Psychology, 44,* 59-127.

Gawronski, B., & Bodenhausen, G. V. (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. (2017). Beyond persons and situations: An interactionist approach to understanding implicit bias. *Psychological Inquiry, 28,* 268-272.

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

Gawronski, B., Cunningham, W. A., LeBel, E. P., & Deutsch, R. (2010). Attentional influences on affective priming: Does categorization influence spontaneous evaluations of multiply categorizable objects? *Cognition and Emotion, 24,* 1008-1025.

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

Gawronski, B., Deutsch, R., & Banse, R. (2011). Response interference tasks as indirect measures of automatic associations. In K. C. Klauer, A. Voss, & C. Stahl (Eds.), *Cognitive methods in social psychology* (pp. 78-123). New York: Guilford Press.

Gawronski, B., Deutsch, R., LeBel, E. P., & Peters, K. R. (2008). Response interference as a mechanism underlying implicit measures: Some traps and gaps in the assessment of mental associations with experimental paradigms. *European Journal of Psychological Assessment, 24,* 218-225.

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,* 573-589.

Gawronski, B., Hofmann, W., & Wilbur, C. J. (2006). Are "implicit" attitudes unconscious? *Consciousness and Cognition, 15,* 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, 2,* 181-193.

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

Gawronski, B., & Payne, B. K. (Eds.). (2010). *Handbook of implicit social cognition: Measurement, theory, and applications*. New York: Guilford Press.

Gawronski, B., Peters, K. R., Brochu, P. M., & Strack, F. (2008). Understanding the relations between different forms of racial prejudice: A cognitive consistency perspective. *Personality and Social Psychology Bulletin, 34,* 648-665.

Gawronski, B., Rydell, R. J., De Houwer, J., Brannon, S. M., Ye, Y., Vervliet, B., & Hu, X. (2018). Contextualized attitude change. *Advances in Experimental Social Psychology, 57,* 1-52.

Gawronski, B., Rydell, R. J., Vervliet, B., & De Houwer, J. (2010). Generalization versus contextualization in automatic evaluation. *Journal of Experimental Psychology: General, 139,* 683-701.

Gawronski, B., & Sritharan, R. (2010). Formation, change, and contextualization of mental associations: Determinants and principles of variations in implicit measures. In B. Gawronski, & B. K. Payne (Eds.), *Handbook of implicit social cognition: Measurement, theory, and applications* (pp. 216-240). New York: Guilford Press.

Gawronski, B., & Strack, F. (2004). On the propositional nature of cognitive consistency: Dissonance changes explicit, but not implicit attitudes. *Journal of Experimental Social Psychology, 40,* 535-542.

Gawronski, B., & Ye, Y. (2014). What drives priming effects in the affect misattribution procedure? *Personality and Social Psychology Bulletin, 40,* 3-15.

Gawronski, B., Ye, Y., Rydell, R. J., & De Houwer, J. (2014). Formation, representation, and activation of contextualized attitudes. *Journal of Experimental Social Psychology, 54,* 188-203.

Govan, C. L., & Williams, K. D. (2004). Changing the affective valence of the stimulus items influences the IAT by re-defining the category labels. *Journal of Experimental Social Psychology, 40,* 357-365.

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

Greenwald, A. G., Banaji, M. R., & Nosek, B. A. (2015). Statistically small effects of the Implicit Association Test can have societally large effects. *Journal of Personality and Social Psychology, 108,* 553-561.

Greenwald, A. 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,* 1464-1480.

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,* 17-41.

Gschwendner, T., Hofmann, W., & Schmitt, M. (2008). Differential stability: The effects of acute and chronic construct accessibility on the temporal stability of the Implicit Association Test. *Journal of Individual Differences, 29,* 70-79.

Hahn, A., & Gawronski, B. (in press). Facing one's implicit biases: From awareness to acknowledgement. *Journal of Personality and Social Psychology*.

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

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 measure. *Personality and Social Psychology Bulletin, 31,* 1369-1385.

Hofmann, W., Gschwendner, T., Castelli, L., & Schmitt, M. (2008). Implicit and explicit attitudes and interracial interaction: The moderating role of situationally available control resources. *Group Processes and Intergroup Relations, 11,* 69-87.

Hofmann, W., Gschwendner, T., Friese, M., Wiers, R., & Schmitt, M. (2008). Working memory capacity and self-regulatory behavior: Towards and individual differences perspective on behavior determination by automatic versus controlled processes. *Journal of Personality and Social Psychology, 95,* 962-977.

Hofmann, W., Gschwendner, T., Nosek, B. A., & Schmitt, M. (2005). What moderates implicit-explicit consistency? *European Review of Social Psychology, 16,* 335-390.

Hofmann, W., Rauch, W., & Gawronski, B. (2007). And deplete us not into temptation: Automatic attitudes, dietary restraint, and self-regulatory resources as determinants of eating behavior. *Journal of Experimental Social Psychology, 43,* 497-504.

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

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

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

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

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

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

Kühnen, U., Schiessl, M., Bauer, N., Paulig, N., Pöhlmann, C., & Schmidthals, K. (2001). How robust is the IAT? Measuring and manipulating implicit attitudes of East- and West-Germans. *Zeitschrift für Experimentelle Psychologie, 48,* 135-144.

Kurdi, B., Seitchik, A. E., Axt, J. R., Carroll, T. J., Karapetyan, A., Kaushik, N., Tomezsko, D., Greenwald, A. G., & Banaji, M. R. (in press). Relationship between the Implicit Association Test and intergroup behavior: A meta-analysis. *American Psychologist*.

Lai, C. K., Marini, M., Lehr, S. A., Cerruti, C., Shin, J. 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,* 1765-1785.

Lai, C. K., Skinner, A. L., Cooley, E., Murrar, S., Brauer, M., Devos, T., … & Nosek, B. A. (2016). Reducing implicit racial preferences: II. Intervention effectiveness across time. *Journal of Experimental Psychology: General, 145,* 1001-1016.

Ledgerwood, A., Eastwick, P. W., & Smith, L. K. (in press). Toward an integrative framework for studying human evaluation: Attitudes toward objects and attributes. *Personality and Social Psychology Review*.

Leippe, M. R., & Eisenstadt, D. (1994). Generalization of dissonance reduction: Decreasing prejudice through induced compliance. *Journal of Personality and Social Psychology, 67,* 395-413.

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

Livingston, R. W. (2002). The role of perceived negativity in the moderation of African Americans' implicit and explicit racial attitudes. *Journal of Experimental Social Psychology, 38,* 405-413.

Lord, C. G., & Lepper, M. R. (1999). Attitude representation theory. *Advances in Experimental Social Psychology, 31,* 265-343.

Ma, D. S., Correll, J., & Wittenbrink, B. (2016). Context dependency at recall: Decoupling context and targets at encoding. *Social Cognition, 34,* 119-132.

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

Maddux, W. W., Barden, J., Brewer, M. B., & Petty, R. E. (2005). Saying no to negativity: The effects of context and motivation to control prejudice on automatic evaluative responses. *Journal of Experimental Social Psychology, 41,* 19-35.

McConahay, J. B. (1986). Modern racism, ambivalence, and the Modern Racism Scale. In J. F. Dovidio, & S. L. Gaertner (Eds.), *Prejudice, discrimination, and racism* (pp. 91-126). New York: Academic Press.

Mischel, W., & Shoda, Y. (1995). A cognitive-affective system theory of personality: Reconceptualizing situations, dispositions, dynamics, and invariance in personality structure. *Psychological Review, 102,* 246-268.

Mitchell, G. (2018). An implicit bias primer. *Virginia Journal of Social Policy and the Law, 25,* 27-59.

Mitchell, J. P., Nosek, B. A., & Banaji, M. R. (2003). Contextual variations in implicit evaluation. *Journal of Experimental Psychology: General, 132,* 455-469.

Molden, D. (Ed.). (2014). *Understanding priming effects in social psychology*. New York: Guilford Press.

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,* 395-417.

Nosek, B. A., & Smyth, F. L. (2007). A multitrait-multimethod validation of the Implicit Association Test: Implicit and explicit attitudes are related but distinct constructs. *Experimental Psychology, 54,* 14-29.

Olson, M. A., & Fazio, R. H. (2003). Relations between implicit measures of prejudice: What are we measuring? *Psychological Science, 14,* 636-639.

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

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

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,* 277-293.

Payne, B. K., & Gawronski, B. (2010). A history of implicit social cognition: Where is it coming from? Where is it now? Where is it going? In B. Gawronski, & B. K. Payne (Eds.), *Handbook of implicit social cognition: Measurement, theory, and applications* (pp. 1-15). New York: Guilford Press.

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

Perugini, M., Richetin, J., & Zogmeister, C. (2010). Prediction of behavior. In B. Gawronski, & B. K. Payne (Eds.), *Handbook of implicit social cognition: Measurement, theory, and applications* (pp. 255-277). New York: Guilford Press.

Peters, K. R., & Gawronski, B. (2011). Are we puppets on a string? Comparing the impact of contingency and validity on implicit and explicit evaluations. *Personality and Social Psychology Bulletin, 37,* 557-569.

Petty, R. E., Fazio, R. H., & Briñol, P. (2009). The new implicit measures: An overview. In R. E. Petty, R. H. Fazio, & P. Briñol (Eds.). *Attitudes: Insights from the new implicit measures* (pp. 3-18). New York: Psychology Press.

Prestwich, A., Perugini, M., Hurling, R., & Richetin, J. (2010). Using the self to change implicit attitudes. *European Journal of Social Psychology, 40,* 61-71.

Rae, J. R., & Olson, K. R. (2018). Test-retest reliability and predictive validity of the Implicit Association Test in children. *Developmental Psychology, 54,* 308-330.

Richeson, J. A., & Ambady, N. (2001). Who's in charge? Effects of situational roles on automatic gender bias. *Sex Roles, 44,* 493-512.

Richeson, J. A., & Ambady, N. (2003). Effects of situational power on automatic racial prejudice. *Journal of Experimental Social Psychology, 39,* 177-183.

Richetin, J., Perugini, M., Adjali, I., & Hurling, R. (2007). The moderator role of intuitive versus deliberative decision making for the predictive validity of implicit and explicit measures. *European Journal of Personality, 21,* 529-546.

Rudman, L. A., Greenwald, A. G., Mellott, D. S., & Schwartz, J. L. K. (1999). Measuring the automatic components of prejudice: Flexibility and generality of the Implicit Association Test. *Social Cognition, 17,* 437-465.

Rudman, L. A., Phelan, J. E., & Heppen, J. B. (2007). Developmental sources of implicit attitudes.

Personality and Social Psychology Bulletin, 33, 1700-1713.

Rydell, R. J., & Gawronski, B. (2009). I like you, I like you not: Understanding the formation of context-dependent automatic attitudes. *Cognition and Emotion, 23,* 1118-1152.

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

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

Sherman, J. W., Gawronski, B., Gonsalkorale, K., Hugenberg, K., Allen, T. J., & Groom, C. J. (2008). The self-regulation of automatic associations and behavioral impulses. *Psychological Review, 115,* 314-335.

Sherman, J. W., Klauer, K. C., & Allen, T. J. (2010). Mathematical modeling of implicit social cognition: The machine in the ghost. In B. Gawronski & B. K. Payne (Eds.), *Handbook of implicit social cognition: Measurement, theory, and applications* (pp. 156-175). New York: Guilford Press.

Snyder, M., & Ickes, W., (1985). Personality and social behavior. In G. Lindzey & (Eds.), *Handbook of social psychology* (3rd ed.; Vol. 2, pp. 248-305). New York: Random House.

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

Wagenmakers, E.-J., Beek, T., Dijkhoff, L., Gronau, Q. F., Acosta, A., Adams, R. B., Jr., … & Zwaan, R. A. (2016). Registered replication report: Strack, Martin, & Stepper (1988). *Perspectives on Psychological Science, 11,* 917-928.

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.

Whitten, S. (2018, July 3). NAAC to Starbucks: Don't let up on weeding out racial bias. *USA Today*. Retrieved from https://www.usatoday.com.

Wilson, T. D., Dunn, D. S., Kraft, D., & Lisle, D. J. (1989). Introspection, attitude change, and attitude-behavior consistency: The disruptive effects of explaining why we feel the way we do. *Advances in Experimental Social Psychology, 22,* 287-343.

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

Wittenbrink, B. Judd, C. M., & Park, B. (2001). Spontaneous prejudice in context: Variability in automatically activated attitudes. *Journal of Personality and Social Psychology, 81,* 815-827.

Ye, Y., Tong, Y.-Y., Chiu, C.-Y., & Gawronski, B. (2017). Attention to context during evaluative learning and context-dependent automatic evaluation: A cross-cultural analysis. *Journal of Experimental Social Psychology, 70,* 1-7.

Table 1. Normative implications of the six lessons for a cogent science of implicit bias.

Lesson 1: Awareness
- Specify which aspect of implicit bias is assumed to be outside of awareness (i.e., source, content, impact).
- Specify whether unawareness of this aspect is assumed to be unique to implicit bias.
- Provide empirical evidence for any hypotheses about unawareness.
- If no evidence can be provided, refrain from making claims about unawareness or explicitly describe them as speculative.
Lesson 2: Conceptual Correspondence
- Avoid confounds between type of measure (implicit vs. explicit) and different contents (e.g., exemplars vs. categories).
- If there is no conceptual correspondence, discuss alternative interpretations of dissociations in terms of different contents.
Lesson 3: Relations to Behavior
- Ensure conceptual correspondence between predictor measures and behavioral criteria.
- Test moderators of predictive relations, including type of behavior, conditions of behavior, and individual differences.
- Ensure comparable reliabilities for different predictor measures as well as behavioral criteria.
Lesson 4: Temporal Stability
- Consider that low temporal stability of implicit bias can be detrimental to prediction of behavior over time.
- Consider that changes in implicit bias scores may reflect either stable changes in traits or transient changes in states.
Lesson 5: Context Effects
- Aim for consistency in measurement contexts in studies on prediction of behavior over time.
- To investigate effectiveness of bias interventions, include follow-up measurements and measurements in different contexts.
Lesson 6: Lack of Process-Purity
- Analyze data with formal modeling procedures to disentangle contributions of multiple distinct processes.
- Replicate findings with implicit measures that are based on different underlying mechanisms.

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 scales 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.