RUNNING HEAD: BIAS IN IMPLICIT MEASURES AS BEHAVIOR

# Bias in Implicit Measures as Instances of Biased Behavior under Suboptimal Conditions in the Laboratory

Jan De Houwer and Yannick Boddez

Ghent University, Belgium

*In Press. Psychological Inquiry*.

mailing address: Jan De Houwer
Ghent University
Henri Dunantlaan 2
B-9000 Ghent
Belgium
email: Jan.DeHouwer@UGent.be
phone: 0032 9 264 64 45

**Abstract**

In this commentary, we note our agreement with many of the statements made by Gawronski et al. (this issue), in particular the idea that implicit bias (IB) is a behavioral phenomenon that can be observed both in the laboratory (e.g., bias in implicit measures; BIM) as well as outside of the laboratory. We also discuss two points of disagreement. First, we argue that there is merit in using the concept *implicit* as an umbrella term that covers several automaticity features. Each automaticity feature refers to a different way in which conditions for cognitive processing are suboptimal (e.g., lack of awareness, lack of motivation). From this perspective, there is merit in contrasting bias that occurs under optimal conditions not only with unconscious bias but also with bias that occurs under other suboptimal conditions (e.g., unintentional bias). Second, we argue that BIM can offer both an educational tool and a laboratory model for IB in the outside world. We discuss how these uses of BIM can be optimized.

The target paper of Gawronski, Ledgerwood, and Eastwick (this issue) provides a valuable contribution to the literature on implicit bias (IB). We find ourselves in agreement with many of the points that the authors put forward. Most importantly, we agree that it is important to realize that scores on implicit measurement tasks such as the Implicit Association Test (IAT) cannot by default be interpreted as instances of unconscious bias. We also agree that the focus on bias in implicit measures (BIM) may have slowed progress in research on IB, that the focus of bias research should be on reducing real-world instances of bias, and that societal disparities can result in social discrimination in a way that is not captured by the psychological concept of bias. We are happy to see that Gawronski et al. share many aspects of our perspective on IB and implicit measures (see De Houwer, 2006, 2014, 2019; De Houwer et al., 2009, 2013, 2021). Most importantly, (a) IB can indeed be conceived of as a behavioral phenomenon that refers to the impact of social cues on behavior, and (b) implicit measures are not the same as indirect measures, nor do they necessarily reflect associative processes. In sum, we support much of what Gawronski et al. put forward in their target paper.

Nevertheless, we also disagree with Gawronski et al. (this issue) on some points. First, we continue to believe IB should not be limited to unconscious bias but should include also instances of bias that are automatic in other ways (e.g., unintentional). Second, we continue to see a potential role for BIM in research on IB, more specifically as an educational tool and as a lab model of IB in the real world.

## Implicit as an Umbrella Term for Suboptimal Processing Conditions

In most of the paper, Gawronski et al. (this issue) use the term *implicit* in the sense of *unconscious*. In the section on the meaning of the term *implicit*, however, they correctly point

out that *implicit* could also be used as a synonym for *automatic* (see De Houwer, 2006; De Houwer et al., 2009). From a decompositional perspective, the concept *automatic* encompasses several non-overlapping automaticity features such as unintentional, fast, efficient, and unconscious (Bargh, 1992; Moors & De Houwer, 2006). Although some colleagues have argued that there is little merit in continuing to use the concepts *implicit* and *automatic* as umbrella terms that cover several automaticity features (e.g., Corneille & Hütter, 2020; Fiedler & Hütter, 2014; Moors, 2016), we do see a need for these concepts. During much of the history of psychology, scientists have examined thinking and behavior under optimal conditions. This is akin to studying the peak performance of a system (e.g., a car). However, there is also merit in studying the performance of a system when it is under pressure, that is, when conditions are suboptimal in some way (De Houwer et al., 2021, Box 2). The features that are typically put under the umbrella of the term *automatic* can be conceptualized as conditions that are suboptimal for cognitive processing (Moors, 2016). The feature *fast* implies a lack of time. *Unconscious* refers to a lack of awareness. *Efficient* refers to a lack of cognitive resources or the presence of demanding other tasks. *Unintentional* refers to a lack of motivation. When bias is defined as the impact of social cues on behavior, IB can thus be understood as the impact of social cues on behavior under suboptimal conditions (De Houwer, 2019; De Houwer et al., 2021). It is intuitively plausible that biased behavior is different (e.g., is less extreme or occurs less frequently) under optimal conditions than under suboptimal conditions. These differences might well depend on the exact way in which conditions are suboptimal (e.g., lack of time vs. lack of awareness). Hence, there is potential merit in studying bias under various suboptimal conditions rather than focusing only on the contrast between, on the one hand, bias under optimal conditions and, on the other hand, bias in the absence of awareness.

In line with what Gawronski et al. (this issue) argue, scores on implicit measurement tasks such as the IAT can be regarded as instances of biased behavior. For instance, the speed and accuracy of responding to the stimuli in the IAT task is likely to be a function of the social cues (e.g., whether a picture shows a person with a light or dark skin). In our opinion, scores on implicit measurement tasks also qualify as instances of *implicit* bias in that the impact of the social cues on performance occurs under suboptimal conditions (e.g., when there is little time to respond; De Houwer, 2019). This point deserves emphasis: From a behavioral perspective, BIM is not a measure or proxy of some other thing that we call IB; rather, it is an instance of IB. Whereas many instances of IB occur in the real-world, BIM is an instance of IB in the laboratory.

Just like different instances of IB outside of the laboratory can differ with regard to the way in which conditions are suboptimal, also different instances of IB in the laboratory (e.g., BIM) can differ with regard to how conditions are suboptimal. We agree with Gawronski et al. (this issue) that current examples of BIM (e.g., scores on IAT tasks) are unlikely to qualify as instances of unconscious bias and are therefore not suitable for the study of unconscious bias. However, those same examples of BIM are likely to be instance of unintentional bias and are therefore potentially informative for the study of unintentional bias. Note that it is also important to verify whether instances of IB outside of the laboratory qualify as instances of unconscious or unintentional bias. BIM and IB outside of the laboratory can be considered as instances of the same type of IB provided that there is evidence showing that both involve bias under the same suboptimal conditions. In past research on IB, however, little effort has been invested in determining whether instances of BIM and IB in the real world match in terms of the way in which conditions are suboptimal for cognitive processing.

## Why BIM Remains Useful for the Study of IB

Once it has been established that specific instances of BIM and IB in the outside world are examples of the same type of IB, there are several ways in which those instances of BIM can be useful for the study of that type of IB. First, BIM can serve as an educational tool. The lab environment is ideally suited for making causal inferences, including the inference that behavior has been influenced by social features of stimuli. It also allows for the best possible control over the way in which conditions are suboptimal. Hence, BIM can provide very convincing evidence for specific types of IB, evidence that can be used to educate people about the phenomenon of IB. Moreover, when completing implicit measurement tasks, people sometimes subjectively experience that social stimulus features influence their responses. This subjective experience could have a bigger impact on the way people think about IB than scientifically more convincing evidence from well-controlled lab studies. Regardless of what aspects of BIM (objective or subjective) are educationally most impactful, from a behavioral perspective, education about IB does not require debates about the normative implications of IB, the mental mechanisms that mediate IB, or the validity and reliability of BIM as a measure of individual differences in IB. Instead, a behavioral perspective helps us to focus on what arguably lies at the core of IB: the fact that social aspects of the environment can influence behavior under suboptimal conditions (De Houwer, 2019; De Houwer et al., 2021).

A second potential reason to cherish BIM as instances of IB in the laboratory is that BIM can provide a lab model of IB. This implies that knowledge generated about BIM in the lab (e.g., which variables moderate BIM; which interventions change BIM; which individuals show large BIM) can be used to increase knowledge about instances of IB outside of the lab.

Especially when it is difficult to study IB outside of the lab, it would be useful to have good lab models of IB. There have been extensive debates about the merits of lab models of real-world phenomena, debates that are often framed in terms of the external validity of experimental research (e.g., Cesario, in press; van den Hout et al., 2017). Our take on this issue is that external validity corresponds to functional similarity, that is, the extent to which variables that affect the model phenomenon in the lab also affect the target phenomenon outside of the lab (e.g., De Houwer, 2020; Vervliet & Boddez, 2020; also see Hesse, 1963). For instance, classical conditioning of fear responses in the lab is a useful model of phobia in the real-world because many of the moderators of fear conditioning that have been discovered in the lab (e.g., the impact of presenting the conditioned stimulus on its own or the impact of contextual cues) seem to also moderate phobias in the real-world. Hence, psychologists can explore the properties of fear conditioning in the lab with the hope of extrapolating that knowledge to phobias in the real world. Likewise, a good lab model of IB would allow us to predict differences in IB in the real world by observing differences in IB in the lab, as well as to influence IB in the real world by using interventions that also change IB in the lab. Of course, no two phenomena are perfectly equivalent. Hence, one should expect to find some differences between the moderators of a lab-based phenomenon and a real-world event. Moreover, because different mechanisms can in principle produce similar phenomena, the usefulness of lab models does not require the assumption that the mechanisms underlying the lab-based phenomenon and the real-world phenomenon are identical. What really counts for practical purposes is to have some degree of functional similarity (also see De Houwer, 2021).

If the usefulness of lab models is primarily an issue of the degree of functional similarity, then the trick is to identify lab-based phenomena that have a high degree of

functional similarity with the real-world phenomenon that one wishes to model. A first strategy to maximize functional similarity is by maximizing phenomenological (i.e., topographical) similarity between the lab-based and real-world situations. This can, for instance, be achieved by using virtual reality tools to recreate real-world situations in the lab (e.g., a driving simulator; see De Houwer et al., 2021, p. 837). The assumption behind this strategy is that situations that look the same also produce behavior that is functionally the same. Although this assumption might often hold, it is important to realize that phenomenological similarity is neither sufficient nor necessary for functional similarity: things that look very similar (e.g., a zebra and a horse) might respond very different to the same interventions (e.g., attempts to ride the animal) whereas things that look more different (e.g., an ostrich and a horse) might respond in similar ways. Moreover, in some cases it might be difficult to create in the lab a situation that is phenomenologically similar to a real-world situation one is interested in.

A second strategy is to select a lab-based phenomenon that is mediated by similar mental mechanisms as a real-world phenomenon. For instance, it could be argued that BIM is mediated by the same attitude representations as IB in the real world (e.g., a negative attitude toward black people). It is, however, notoriously difficult to reach consensus about the nature of mental representations and the mental processes via which these representations influence behavior (e.g., De Houwer et al., 2013), including IB (De Houwer et al., 2021).

A third strategy is to select lab-based situations in which behavior is known to be functionally similar to a real-world behavior in at least some respects. The assumption here is that lab-based and real-world phenomena that are known to be functionally similar in some respects are likely to be functionally similar also in other respects. Also this assumption

might not always hold, however. Moreover, applying this strategy also requires pre-existing knowledge of the functional properties of behavior in the lab and in real-world situations, knowledge that might not be available.

Although none of the three strategies guarantees success in building useful lab-based models of real world phenomena, their use is likely to increase the chances of success. Based on this consideration, it is unsurprising that currently, BIM provides a poor lab model of IB in the real world. Phenomenologically, implicit measurement procedures (e.g., an IAT task) are often very different from real-world situations in which IB occurs. In fact, many implicit measurement tasks (e.g., evaluative priming, see Fazio et al., 1995; affective Simon tasks, see De Houwer, 2003) were modelled not after instances of IB in the real world but after tasks that were developed by cognitive psychologists for use in the lab (e.g., the priming or Simon task). In terms of mental mechanisms, there is little agreement on the nature of the mental representations that mediate IB (e.g., are it associations or propositions; see De Houwer et al., 2021) or the mental processes via which those representations produce IB (e.g., biased interpretation; see Gawronski et al., this issue). Hence, little is known about the extent to which similar mental mechanisms mediate BIM and IB outside of the laboratory. In terms of functional similarity, little attention has been given to determining the functional properties of instances of IB (e.g., whether it depends on the nature of the social cues, such as faces vs. names), either in the lab or in the real-world. Hence, there is little knowledge about functional differences between BIM and IB in the real-world.

Although it is thus understandable that current examples of BIM are relatively poor lab models of IB in the real world, in principle, more useful lab models of IB could be developed in the future. Virtual reality tools certainly offer a potential way forward. Like

Gawronski et al. (this issue), we also encourage researchers to redirect their attention to real-world instances of IB. Learning more about IB in the real world is an essential step toward developing lab-based models of IB. In studying real-world instances of IB, attention should be directed not only at uncovering the mental mechanisms that mediate IB in the real world but also at documenting the variables that moderate IB in the real world. This information can then be used to improve the external validity of BIM as a lab model of IB in the real world, which implies that BIM will become a more useful tool for predicting and influencing IB in the real world. These future instances of lab-based IB might have little in common with what we now call BIM. Such an evolution is perfectly acceptable when scores on implicit measures are no longer thought of as proxies of unobservable entities that guide behavior but rather as indices of behavior in the lab that can inform us about behavior outside of the lab.

## References

Bargh, J. A. (1992). The ecology of automaticity: Toward establishing the conditions needed to produce automatic processing effects. *The American Journal of Psychology, 105*, 181–199.

Cesario, J. (in press). What can experimental studies of bias tell us about group disparities? *Behavioral and Brain Sciences*.

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

De Houwer, J. (2003). The extrinsic affective Simon task. *Experimental Psychology, 50*, 77-85.

De Houwer, J. (2006). What are implicit measures and why are we using them. In R. W. Wiers & A. W. Stacy (Eds.), *The handbook of implicit cognition and addiction* (pp. 11-28). Thousand Oaks, CA: Sage Publishers.

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

De Houwer, J. (2019). Implicit bias is behavior: A functional-cognitive perspective on implicit bias. *Perspectives on Psychological Science, 14*, 835-840.

De Houwer, J. (2020). Revisiting classical conditioning as a model for anxiety disorders: A conceptual analysis and brief review. *Behaviour Research and Therapy, 127*, 103558.

De Houwer, J. (2021). On the challenges of cognitive psychopathology research and possible ways forward: Arguments for a pragmatic cognitive approach. *Current Opinion in Psychology, 41*, 96-99.

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

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

De Houwer, J., Van Dessel, P., & Moran, T. (2021). Attitudes as propositional representations. *Trends in Cognitive Sciences, 25*, 870-882. https://doi.org/10.1016/j.tics.2021.07.003

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.

Fiedler, K., & Hütter, M. (2014). The limits of automaticity. In J. Sherman, B. Gawronski, & Y. Trope (Eds.), *Dual process theories of the social mind* (pp. 497-513). New York: Guilford Press.

Gawronski, B., Ledgerwood, A., & Eastwick, P. W. (this issue). Implicit Bias ≠ Bias on Implicit Measures. *Psychological Inquiry*.

Hesse, M. (1963). *Models and Analogies in Science*. London: Sheed and Ward.

Moors, A. (2016). Automaticity: Componential, causal, and mechanistic explanations. *Annual Review of Psychology, 67*, 263–287.

Moors, A., & De Houwer, J. (2006). Automaticity: A conceptual and theoretical analysis. *Psychological Bulletin, 132*, 297–326.

van den Hout, M. A., Engelhard, I. M., & McNally, R. J. (2017). Thoughts on experimental psychopathology. *Psychopathology Review*, *4*, 141-154.

Vervliet, B., & Boddez, Y. (2020). Memories of 100 years of human fear conditioning research and expectations for its future. *Behaviour Research and Therapy*, 103732.

## Acknowledgements

Jan De Houwer, Yannick Boddez, Ghent University, Ghent, Belgium. Correspondence should be addressed to Jan De Houwer, Ghent University, Henri Dunantlaan 2, B-9000 Ghent, Belgium. Electronic mail can be sent to Jan.DeHouwer@UGent.be . The preparation of this paper was made possible by Ghent University Grant BOF16/MET_V/002 to Jan De Houwer.

## Funding Information

The preparation of this paper was made possible by Ghent University Grant BOF16/MET_V/002 to Jan De Houwer.

## Declaration of interest: none