Behavioral and Brain Sciences

cambridge.org/bbs

What can experimental studies of bias tell us about real-world group disparities?

Joseph Cesario

Department of Psychology, Michigan State University, East Lansing, MI 48824, USA.
cesario@msu.edu
www.cesariolab.com

Target Article

Cite this article: Cesario J. (2022) What can experimental studies of bias tell us about real-world group disparities? *Behavioral and Brain Sciences* **45**, e66: 1–71. doi:10.1017/S0140525X21000017

Target Article Accepted: 1 January 2021
Target Article Manuscript Online: 8 January 2021
Commentaries Accepted: 1 May 2021

Key words:
Discrimination; disparate outcomes; implicit bias; racial bias; school discipline; shooter bias; social psychology; stereotyping

Abstract

This article questions the widespread use of experimental social psychology to understand real-world group disparities. Standard experimental practice is to design studies in which participants make judgments of targets who vary only on the social categories to which they belong. This is typically done under simplified decision landscapes and with untrained decision-makers. For example, to understand racial disparities in police shootings, researchers show pictures of armed and unarmed Black and White men to undergraduates and have them press "shoot" and "don't shoot" buttons. Having demonstrated categorical bias under these conditions, researchers then use such findings to claim that real-world disparities are also due to decision-maker bias. I describe three flaws inherent in this approach, flaws which undermine any direct contribution of experimental studies to explaining group disparities. First, the decision landscapes used in experimental studies lack crucial components present in actual decisions (*missing information flaw*). Second, categorical effects in experimental studies are not interpreted in light of other effects on outcomes, including behavioral differences across groups (*missing forces flaw*). Third, there is no systematic testing of whether the contingencies required to produce experimental effects are present in real-world decisions (*missing contingencies flaw*). I apply this analysis to three research topics to illustrate the scope of the problem. I discuss how this research tradition has skewed our understanding of the human mind within and beyond the discipline and how results from experimental studies of bias are generally misunderstood. I conclude by arguing that the current research tradition should be abandoned.

1. Introduction

For more than half a century, experimental social psychologists have (1) demonstrated the many ways people are treated differently because of their race, age, sex, and other social categories and (2) used these findings to explain why group disparities exist in the real world. From racial disparities in fatal police shootings and school discipline, to sex disparities in science, technology, engineering, and mathematics (STEM) engagement and corporate leadership, social psychologists have overwhelmingly concluded that the stereotypes in the heads of decision-makers play a substantial role in causing group disparities, whether or not people agree with or even consciously acknowledge such stereotypes (Devine, 1989; Greenwald & Banaji, 1995). The logic among social psychologists has been the following: If we can show in an experiment that people are treated differently based on their outward appearances when we present them as equal in all other respects, then in the real world such differential treatment exists and is a major cause for why outcomes differ across groups (see, e.g., Greenwald & Krieger, 2006; Kang & Banaji, 2006). As just one example illustrating the way in which experimental demonstrations of decision-maker bias have been tied to disparate outcomes, Moss-Racusin, Dovidio, Brescoll, Graham, and Handelsman (2012) state clearly that their research "informs the debate on possible causes of the gender disparity in academic science by providing unique experimental evidence that science faculty of both genders exhibit bias against female undergraduates" (p. 16477).

The purpose of this article is to show that standard practice in experimental social psychology is fundamentally flawed, so much so that findings from these studies cannot be used to draw any substantive conclusions about the nature of real-world disparities – despite the ubiquitous practice of drawing exactly these conclusions. There are three problems inherent in the current approach that render it impotent for this purpose. First, critical pieces of information used by actual decision-makers are absent in experimental studies (*missing information flaw*). Second, effects of biased decision-making are rarely understood in the context of other important influences on group outcomes, such as the behaviors of targets themselves (*missing forces flaw*). Third, there is no systematic study on whether the contingencies required to produce experimental bias are present in actual decisions (*missing contingencies flaw*). These three flaws can lead researchers to vastly overestimate the role of stereotyping as a causal process,

CAMBRIDGE UNIVERSITY PRESS

even going so far as to reveal experimental stereotyping effects when they play no role in real decisions or in causing group disparities. Although current experimental studies *can* provide important information about stereotyping processes per se, they cannot and do not provide information about the nature of group disparities. That is, the contribution of stereotyping and bias research is misunderstood and misused.

I first describe the standard "research cycle" of stereotyping and bias studies in experimental social psychology. I describe the flaws inherent in this approach at the abstract level and then apply the analysis to three research topics in social psychology: police officers' decision to use deadly force, implicit bias, and school disciplinary policy. I then describe what experimental studies of bias *can* tell us and how researchers generally misinterpret the nature of such studies. I speculate on the ways this research tradition has skewed the understanding of the human mind that has been exported from our discipline to the culture at large. I then connect this critique to related critiques within psychology and similar problems that have arisen in other fields. In the final section, I chart out an alternative path that might be more effective for studying group disparities.

Throughout this paper, I focus on the familiar social psychological demonstrations of categorical bias: experiments in which participants respond differently to targets from different social categories. Although I focus on studies that posit stereotype activation as the culprit for such differential responding (as this is a long-standing way of understanding such effects, e.g., Duncan, 1976), nearly all of the current analysis is applicable to bias caused by other sources. I focus on experimental social psychology because this area has had a considerable impact on the discussion of group disparities, but this is not to say that similar critiques cannot be leveled against other areas and disciplines. The current critique is also distinct from related critiques of mundane realism or external validity, which I discuss in more detail later. Instead, this critique is about how social psychologists are fundamentally misguided in how they approach the study of group disparities, which distorts the nature of the decision under study and leads to incorrect conclusions about the conditions under which decisions will be more or less biased. Although psychology has no shortage of problems to be addressed (e.g., Srivastava, 2016), I limit my discussion in this paper to the misuse of experimental social psychology in explaining group disparities.

Before getting into the details of the argument, it is important to provide two cautionary notes. First, I am addressing the question of whether decision-maker bias produces group disparities in the immediate outcomes of that decision (and whether experimental social psychology can inform this process). This is seen in the example of a police officer's decision to shoot and racial disparities in being shot by police, or of a search committee's hiring decision and sex disparities in STEM employment. The current analysis does not address or dispute the possibility that decision-maker bias may enter earlier in the chain of events that leads to the decision in question. For example, police officers may show bias in the decision to engage in discretionary stopping of Black citizens or high school teachers may show bias in discouraging female students from pursuing STEM careers.

Second, the current analysis relies substantially on the fact that the distributions of behaviors, personality, character, preferences, abilities, and so on are not equal across different demographic groups (and that this fact is not appropriately considered by experimental social psychologists). I make no claims about the origin of these group differences in terms of the degree to which they are caused by individual decision-makers, "structural" forces beyond individual actions, genetic factors, incentive structures because of government policies, and so on. The point here is not to claim that group differences are inherent to people (although they might very well be) or that there are no broader social influences on human behavior. There may be systematic bias that produces group differences in the distributions of important characteristics. For the purposes of the present argument, the distal causes of group differences are irrelevant because these causes are separable from the question of whether group disparities are because of biased decision-making for specific outcomes. For example, the reasons why men and women differ in their interest in things versus people is a separate question from whether faculty search committees are biased against women in hiring for STEM positions.

On both these points, there is the possibility that bias "earlier" in the causal chain eventually leads to disparities on a later outcome, even while decision-makers show no bias on that later outcome. Of course, claims of "earlier" bias also require evidence, and if the available evidence is merely more of the same demonstrations from experimental social psychology, then these studies suffer from the same flaws described here and are, therefore, not convincing evidence.

2. The standard approach

The standard research cycle begins with an observation that groups differ in their real-world outcomes and the desire to understand the causes of such disparities. Simply, we see that members of some groups get better or worse outcomes than members of other groups and we want to know why. It is, perhaps, natural that social psychologists would start with the assumption that stereotypes – categorical information stored in a decision-maker's mind – play a meaningful role in producing these group differences. To gather evidence in support of this possibility, researchers design experiments in which participants make judgments of targets who vary only with respect to the social categories to which they belong. For example, to study the role of race in police officers' decision to use deadly force, researchers show participants pictures of Black and White men who do not vary in how they are presented in any way other than their race (as in the First-Person Shooter Task [FPST]; Correll, Park, Judd, & Wittenbrink, 2002). If participants shoot unarmed Blacks more than unarmed Whites, one can be sure that the race of the target played a causal role in participants' decisions because the experimenter has presented the groups in an identical way on all other dimensions (such as their posture, the frequency of holding a gun, facial expressions, etc.). Making all groups exactly equal in how they are presented in an experiment allows the researcher to conclude that the decision-maker (and not differences in the behavior of targets themselves) is responsible for biased responses directed at targets from different groups. It would be difficult to overstate the ubiquity of this approach in experimental social psychology; it is the paragon of systematic design and is understood as *the* method for studying the biasing effects of categories.

Joseph Cesario is professor of psychology at Michigan State University and studies social cognitive processes. He is co-founding editor of *Comprehensive Results in Social Psychology*, the field's first peer-review preregistration journal.

https://doi.org/10.1017/S0140525X21000017 Published online by Cambridge University Press

Having established that social categories impact participants' decisions in an experiment, researchers return to the original real-world disparity and conclude that the same processes observed in the lab explain these disparities as well (see, e.g., Moss-Racusin et al., 2012 for a prototypical example). That is, if stereotypes cause people to treat targets differently when there are no real behavioral differences in experimental stimuli, then this same biased treatment on the part of decision-makers is at play in the real world and can account for a meaningful amount of the disparities we see across groups. Researchers then complete the circle by using their experimental findings as evidence for designing interventions intended to reduce the disparity of interest.

3. Critical flaws of the standard paradigm

The standard experimental approach in social psychology contains three fundamental flaws which prevent the findings of experimental studies from being directly applied to the study of group disparities: *the flaw of missing information*, *the flaw of missing forces*, and *the flaw of missing contingencies*. The first flaw is that the decision components used by real-world decision-makers are absent in our experiments; in other words, information that is available to and used by actual decision-makers is removed from our experimental studies. The second flaw is that other influences on group outcomes – such as actual behavioral differences across groups – are not integrated into our designs, analyses, or conclusions. The third flaw is the lack of systematic study of whether the contingencies required to produce experimental bias are present in real decisions; along with this is the understudied question of whether the experimental landscape changes the motivation and ability of decision-makers. By "fatal flaws," I mean that any one of these flaws can reveal experimental stereotyping effects *even when no such effects exist* in real decisions. I first describe these flaws at the general level and then show how such flaws are evident in three different research areas in experimental social psychology. Descriptions and examples of these flaws are summarized in Table 1.

3.1 The missing information flaw

For reasons good and bad, experimental studies of categorical bias in social psychology are massive simplifications of real-world decision landscapes. The problem is that this simplification removes information that may play a strong or even critical role in real decisions. When this happens, three distortions may occur. First, the missing information may have more powerful effects than social category information and may overwhelm any categorical influence in real decisions; when such forces are removed all that remains is the categorical influence, which is then realized in the experiment. Second, removing these variables may leave experimental participants with no useful information to render a judgment other than the target's social category; although categorical information may be used minimally or not at all in real decisions, experimental participants now use it because of the absence of any other kind of diagnostic information. Third, the presence or absence of such information may change the underlying decision process itself, leaving researchers with a distorted understanding of the cognitive dynamics at play in real decisions. In all cases, researchers are at risk of incorrectly concluding that the reliable and replicable effects of categories observed in their experiments are present in the real world and have the same effects on outcomes. This suggests that social psychologists may fundamentally misunderstand the nature of a decision if their experimental methods strip away critical features present in real decision-makers' environments.

This first flaw reflects a fallacy in the justification for using experimental studies, which is to presume that any information that *can* affect outcomes in an experimental setting *does* have the same effect in the real world. Said differently, the fallacy is the unstated belief that adding additional information to the decision landscape will not change the nature or magnitude of an experimental effect and the missing information can therefore safely be ignored.

3.2 The missing forces flaw

The second flaw of the current experimental approach is that researchers do not interpret experimental effects in light of the other causal forces which impact group outcomes in the real world. Primary among these forces is the behavior of the targets themselves and the cognitive, motivational, and behavioral differences that exist across groups. This flaw is important because if there are strong influences on group outcomes besides biased treatment, then it follows that experimental participants may show reliable decision-maker bias – even very strong bias – while such bias exerts no discernable effect on real outcomes. If true, social psychologists may be perpetually disappointed in the state of the world because their recommended interventions of removing decision-maker bias will not yield equal outcomes or even reduce group disparities. Indeed, depending on the strength of group differences, social psychologists may be diverting resources away from effective interventions and toward those that will have little effect on reducing disparities.

This flaw reflects the fallacy that researchers believe they can safely ignore the degree to which the stimuli used in experimental studies match the distributional properties of the real-world groups they represent. One reason for this disregard may be the belief that all groups have roughly identical distributions on important underlying causal characteristics.[1] Yet this assumption is incorrect, as groups differ (and often markedly so) on important personality, motivational, and cognitive dimensions – in other words, on the interest and ability factors that relate to nearly all outcomes (see, e.g., ACT, 2017; Andreoni et al., 2019; Beaver et al., 2013; Benbow & Stanley, 1980; Benbow, Lubinski, Shea, & Eftekhari-Sanjani, 2000; Byrnes, Miller, & Schafer, 1999; Ceci & Williams, 2010; Cesario, Johnson, & Terrill, 2019; Diekman, Steinberg, Brown, Belanger, & Clark, 2017; Gottfredson, 1998; Halpern et al., 2007; Hsia, 1988; Hsin & Xie, 2014; Jussim, Cain, Crawford, Harber, & Cohen, 2009; Jussim, Crawford, Anglin, Chambers, et al., 2015a; Jussim, Crawford, & Rubinstein, 2015c; Lee & Ashton, 2020; Lippa, 1998; Lu, Nisbett, & Morris, 2020; Lubinski & Benbow, 1992; Lynn, 2004; Lynn & Irwing, 2004; McLanahan & Percheski, 2008; Roth, Bevier, Bobko, Switzer, & Tyler, 2001; Sowell, 2005, 2008; Su, Rounds, & Armstrong, 2009; Tregle, Nix, & Alpert, 2019; Wright, Morgan, Coyne, Beaver, & Barnes, 2014).[2] In understanding the role of decision-maker bias in producing disparate outcomes, it is necessary to compare and interpret the size of categorical bias effects with the size of these behavioral differences across groups.

Methodologically, this flaw is guaranteed because target stimuli are presented as equal on all dimensions except for social category membership. Statistically, this flaw is guaranteed because analytic models either do not incorporate information about real-world behavioral differences or if they do, they are treated as

https://doi.org/10.1017/S0140525X21000017 Published online by Cambridge University Press

Table 1. Three flaws inherent to experimental social psychology studies of bias

Flaw	Description	Problem	Fallacy	Example
Flaw of missing information	Experiments strip away critical information used in real-world decisions.	Missing information may overwhelm strength of categorical bias in real decisions. Categorical information may not be used at all when missing information is present. Missing information may change cognitive process of the decision-maker.	Variables that *can* have an effect in an experiment *do* have an effect in real decisions.	Dispatch information is removed from experimental studies of the decision to shoot, although this is present in real decisions; when such information is reintroduced, racial bias is eliminated.
Flaw of missing forces	The strength of experimental demonstrations of categorical bias is not interpreted in light of other influences on group outcomes.	Categorical bias may have no effect on group outcomes relative to other forces. Distorts our understanding of the nature of the disparity.	Assumption that distributions of important outcome-related variables are equal across groups.	Experimental demonstrations of gender bias in STEM are not understood in light of distributional differences in ability and interest related to the outcome.
Flaw of missing contingencies	Failure to appreciate the contingencies required for experimental bias to be realized, some of which can impact the ability and motivation of decision-makers.	Necessary contingencies for bias may not be present in actual decisions. Necessary contingencies for bias may impact decision-makers' ability and motivation in ways that change the expression of bias. Necessary contingencies are not given attention when applying findings to real-world decisions.	The total control of experimental participants does not meaningfully affect applications to real-world decisions.	The parameters of experimental shooter tasks do not match the parameters of real shootings.

control variables whose (often very strong) relationship to the outcome of interest is ignored. These decisions shift researchers' attention away from the role that causal forces beyond categorical bias may have on group disparities in the real world – or at least, allow researchers to relegate these forces to a brief mention in the Introduction of their papers. To the extent that groups differ in important ways, and such differences have strong effects on obtained outcomes, the role of perceiver bias and stereotyping will be overstated.

Although the first flaw concerns removing everything but categorical information from the experiment, this second flaw concerns failing to interpret those experimental categorical effects in light of other known forces on group outcomes. This failure can lead to overemphasizing the role of perceiver bias, as revealed by experimental methods and "statistically significant" model coefficients (while ignoring variance explained or effect sizes).

3.3 The missing contingencies flaw

The third critical flaw is the failure to study whether the precise contingencies needed to produce categorical bias in our experiments are realized in real-world decision situations. Whether the conditions required for experimental demonstrations of bias are present in the real world obviously informs the degree to which such demonstrations can explain group disparities. However, it is also important because such contingencies relate to the motivation and ability of decision-makers in experimental tasks, and it is known that motivation and ability are critical variables for categories to have biasing effects on judgments.

This flaw reflects another fallacy present in experimental studies of bias, which is to ignore the contrived nature of experiments and the total control experimenters exercise over all aspects of the participant's experience. Indeed, contingencies are "missing" in not one but two ways. First, social psychologists do not explore whether the conditions needed for experimental bias are present in the real world. But second, discussion of these conditions is missing when social psychologists advocate for their research outside of academic psychology, where "contingent" and "conditional" bias now becomes "widespread" and "pervasive" bias (e.g., Greenwald & Krieger, 2006; Kang & Banaji, 2006).

Specific contingencies are required for category information to bias a person's decisions; stereotype effects do not occur uniformly for all people or under all conditions. For categories to bias decisions, clear diagnostic or individuating information must be absent and perceivers must lack the ability or motivation to control the biasing influence of categories. When decision-makers have adequate ability and motivation to control the effects of categorical information, or when information is unambiguous (as with strong individuating information or applicability of a single concept; Higgins, 1996), categories have little to no biasing effect on judgments (e.g., Darley & Gross, 1983; Dovidio & Gaertner, 2000; Koch, D'Mello, & Sackett, 2015; Krueger & Rothbart, 1988; Locksley, Borgida, Brekke, & Hepburn, 1980; see Jussim, 2012b; Jussim et al., 2015c; Kunda & Spencer, 2003). As stated unequivocally in a summary by Kunda and Thagard (1996) over two decades ago, "It is clear … that the target's behavior has been shown to undermine the effects of stereotypes based on all the major social categories" (p. 292).

Given the importance of some contingency set, researchers must outline the precise contingencies required to give rise to bias in the lab and *detail the degree to which experimental contingencies are present in real decisions*. Assuming researchers can, in

https://doi.org/10.1017/S0140525X21000017 Published online by Cambridge University Press

fact, show that these necessary contingencies are reproduced with regularity in the real world, researchers are also responsible for keeping these contingencies front and center when discussing their study in applied contexts so as to not overextend claims of bias.

Experimental contingencies are also important because they relate to the roles of ability and motivation in biased decision-making. Ability and motivation have been the twin variables in nearly every major model of impression formation, persuasion, and decision-making in social cognition for decades (Bargh, 1999; Devine, 1989; Fazio, 1990; Fiske & Neuberg, 1990; Petty & Wegener, 1999; Smith & DeCoster, 2000). Given this, it is surprising that social cognitive researchers have not systematically studied whether novice or experimental participants match expert or real-world decision-makers on these two dimensions.

First, regarding ability, it has long been known that experts use different information, and use the same information differently, relative to novices (see, e.g., Klein, 1998; Koch et al., 2015; Levine, Resnick, & Higgins, 1993; Logan, 2018; although not always, see Miller, 2019). In experimental studies of stereotyping, it is undeniable that there are no serious attempts to train participants before having them render a judgment. If trained decision-makers use information in the decision landscape differently than do untrained participants, this represents an important difference in ability between the two groups. If experts attend to different decision components or use these components differently than novices, *and this difference changes the effect of social categories on the ultimate decision*, then the conclusion of widespread bias in real decisions based on findings from undergraduate participants will be unwarranted.

The experimental situation itself can also be understood as impacting participants' ability in important ways. As described above, the simplified experimental methods used in studies of bias remove important sources of information used by real decision-makers. Said differently, researchers change the nature of accuracy and bias in decision-makers when they fail to give participants information that is available in real decisions, information which can allow participants to make decisions in unbiased (or, at least, less biased) ways.

Besides the ability differences between expert decision-makers and naive experimental participants, there are surely important motivational differences as well. Some research has tried to increase participants' motivation to provide unbiased decisions by rewarding accurate decisions or increasing personal relevance, but whether such manipulations produce similar motivation to those found outside experimental contexts is unknown. And importantly, experimental participants simply do not bear the costs of their decisions in ways that are required of many real-world decision-makers, a fact which can change the link between intentions and behavior (e.g., Sowell, 2008; Tetlock, 1985). For naive participants making imaginary decisions about hypothetical targets, there is no effect on their lives once the experiment ends.

3.4 Summary

The three critical flaws of the experimental approach to the study of bias and group disparities can be summarized as follows. If the information used by actual decision-makers in real-world decision landscapes is absent in experimental studies of these decisions, one's understanding of the decision under study can be dramatically skewed. Merely demonstrating bias conveys nothing about the strength of that bias relative to other causal forces on group outcomes. Moreover, there is a failure to specify the required contingencies for experimental demonstrations of bias and explore whether such contingencies are present in real decisions. Finally, if actual decision-makers use information differently or have different motivations and abilities than experimental decision-makers, there is no guarantee that bias will be observed outside experimental contexts. For these reasons, claims of ubiquitous bias among real-world decision-makers may be overstated.

4. Experimental studies of bias: Three topics

Having identified the problems inherent to experimental studies of bias at the general level, I now turn to demonstrating how these problems appear in practice. I chose the three topics discussed next because they cover a range of characteristics. Shooter bias is a narrow topic with nearly two decades of research and is a prototypical social psychological study. Implicit bias is a much broader topic but one that has had a major effect on the public's understanding of group disparities. School disciplinary policies are a relatively new topic, but an important one with broad interest beyond the discipline of psychology.

4.1 Shooter bias

For nearly two decades, researchers have studied the question of racial bias in police officers' decisions to use deadly force. Without question, the most common experimental task used is the FPST, in which participants are shown pictures of armed and unarmed Black or White men and asked to press buttons labeled "shoot" and "don't shoot" (Correll et al., in press; see Cesario & Carrillo, in press for a summary). How does this research fare with respect to the three fundamental flaws of experimental social psychology? (Fig. 1).

4.1.1 Shooter bias: The missing information flaw

With respect to the first flaw, every relevant piece of information used by police officers in the decision to shoot has been removed from the standard experimental task, absent the one variable of whether or not targets are holding guns (an effect which overwhelms all other effects in both real and experimental decisions). Although a small number of exceptions exist and are discussed below, this has been true of virtually all studies using the FPST (see Cesario & Carrillo, in press). These missing variables include: dispatch information about the citizen and why the officer has been called to the scene, neighborhood information, past encounters with the citizen, how the interaction has unfolded leading up to the decision point (e.g., has the citizen been compliant thus far?), the physical movements by the citizen at the moment of the decision point, the goal of the officer at the scene, whether other officers are present, whether non-lethal tactics have already been used, and so on.

Officers report that all these factors matter, and indeed officers are trained to attend to these factors and integrate them into their dynamic, continuously updating decision to use deadly force as the interaction with the citizen unfolds. Of course, whether and the extent to which any of these pieces of information actually affect officers' decisions are empirical questions. Yet by not including these features, researchers simply have no idea whether their experimental methods are adequately capturing officers' decision processes. Researchers may be fundamentally misunderstanding the underlying cognitive decision dynamics if factors

https://doi.org/10.1017/S0140525X21000017 Published online by Cambridge University Press

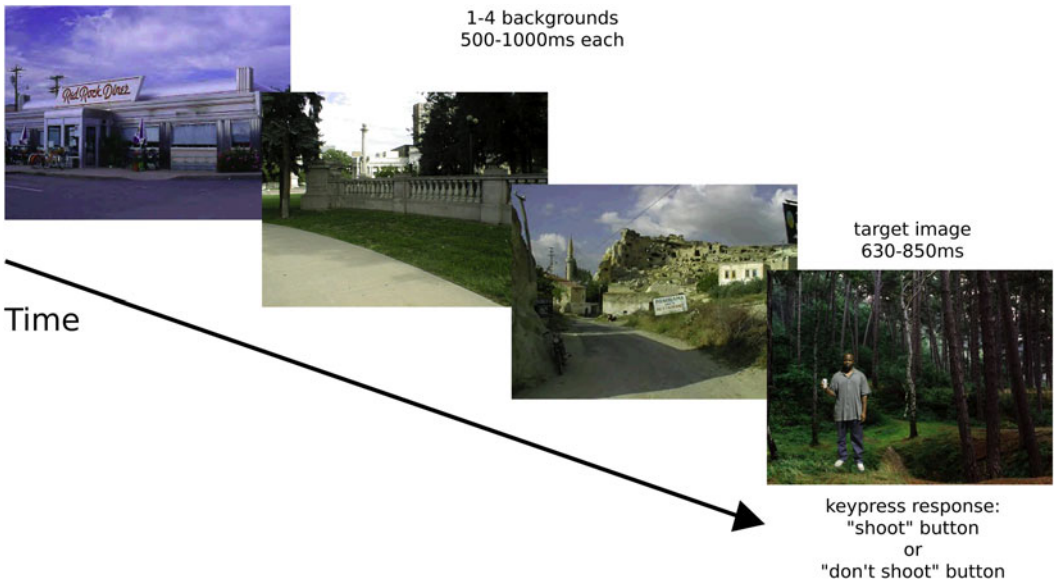

Figure 1. Example trial of the First-Person Shooter Task, the most common experimental method for understanding police officers' decisions to shoot.

that impact those dynamics in real decision-makers have no possibility of impacting experimental participants, simply because researchers have failed to include such factors in their studies. Thus, we can ask, what happens to racial bias – at both the behavioral and the cognitive process levels – when such information is introduced into the experiment?

As one example of how the conclusions from experimental studies can drastically change if we introduce information used by officers in real decisions, Johnson, Cesario, and Pleskac (2018) conducted a series of studies examining the role of dispatch information in the decision to shoot. Participants completed a standard FPST, but with an important modification: On some trials, participants were given dispatch information at the start of each trial that contained the race of the target, whether the target had a weapon (correct 75% of the time), or both pieces of information. As shown in Figure 2, although untrained undergraduates showed the standard race bias effect when no dispatch information was given, dispatch information of any type eliminated race bias in the decision to shoot. Thus, a single change to the standard, simplified experimental task to include the most important and relevant information that officers have in real shootings eliminated the biasing effects of race. This calls into question our ability to draw conclusions about real-world cases of police shootings from simplified experimental paradigms. More generally, it illustrates the importance of ensuring that the decision landscape for participants in experimental laboratory tasks contains those factors used by real-world decision-makers.

This point is consistent with research by Correll and colleagues (Correll, Wittenbrink, Park, Judd, & Goyle, 2011; but see Pleskac, Cesario, & Johnson, 2018), who manipulated the neighborhood background in which targets in the FPST appeared. In nearly all uses of the FPST, targets are presented in neutral, uninformative backgrounds (office buildings, parks, etc.). These researchers manipulated whether targets appeared in neutral backgrounds or dangerous, urban backgrounds. Placing targets in the dangerous backgrounds completely eliminated racial bias in the decision to shoot. To the extent that real-world police shootings occur in dangerous neighborhoods or situations, this seriously calls into question the degree to which our experimental findings inform our understanding of police officer racial bias.

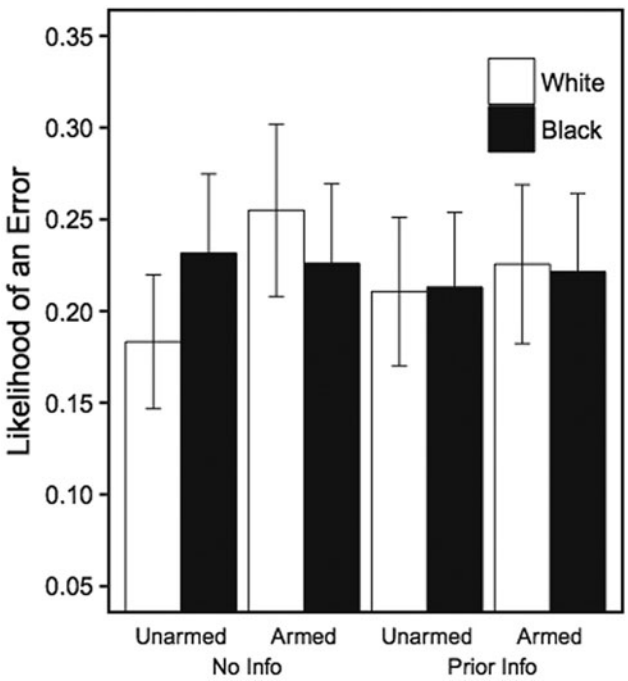

Figure 2. In the standard First-Person Shooter Task ("No Info"), undergraduate participants showed racial bias in the decision to shoot. When provided with prior dispatch information about target race or presence of a weapon ("Prior Info"), participants showed no evidence of racial bias in the decision to shoot. Black and white bars refer to target race. Modified from Johnson et al. (2018).

As another attempt to reintroduce those factors present in actual decisions but missing in experimental studies, at least three independent research groups have used some version of an immersive shooting simulator similar to those used for training by law enforcement (Cox, Devine, Plant, & Schwartz, 2014; James, James, & Vila, 2016; James, Vila, & Daratha, 2013; James,

https://doi.org/10.1017/S0140525X21000017 Published online by Cambridge University Press

Klinger, & Vila, 2014; Pleskac, Johnson, Cesario, Terrill, & Gagnon, under review). As depicted in the right panel of Figure 3, participants in such studies stand in front of a projection screen and watch life-sized videos recorded from a first-person point of view. These videos are of policing scenarios similar to those encountered by law enforcement (e.g., traffic pullovers and domestic disturbances). Participants verbally interact with individuals in the videos as they unfold over time, during which participants must decide whether to use deadly force. This response is made using a modified handgun; when the trigger is pulled, cycling of the firearm occurs through a compressed air connection, which provides recoil and initiates the sound of a handgun firing through a set of speakers. Officers routinely report being highly involved with these scenarios and display strong emotional states, attesting to the realism of the method.

Importantly, such a methodological change is not merely about recreating surface-level similarity to the decision to shoot in terms of the participant's experience (e.g., pressing a button vs. holding a gun). This method allows researchers to introduce back into the decision landscape those factors which simplified tasks remove but which officers report as being important.

Cesario and Carrillo (in press) came to two main conclusions in their summary of the research on shooting simulator studies. First, among the studies that manipulated the scenarios to which officers responded, there was strong evidence of the importance of the scenario and the specific actors on officers' decisions – stronger than the effects of suspect race. In Pleskac et al. (2019) for example, variance in officers' decisions was primarily explained by the different scenarios in the videos (e.g., serving a warrant for armed robbery vs. failure to pay for child support) and the behavior of the different actors in the videos – features that the standard FPST removes entirely from the decision landscape. The second main conclusion was that studies using shooting simulators do *not* provide strong evidence of anti-Black bias in officers' decisions. Indeed, in all possible tests of racial bias across such studies, only about *5%* showed anti-Black bias in officers' decisions. In contrast, almost 40% of tests showed *anti-White* bias in officers' decisions.

Although these results are inconsistent with claims from experimental social psychologists regarding the overwhelming importance of racial stereotypes in decisions to shoot, these results are consistent with the many analyses of actual police shootings that have revealed the importance of context and suspect behavior (see, e.g., Cesario et al., 2019; Fryer, 2016; Fyfe, 1980; Geller & Karales, 1981; Inn, Wheeler, & Sparling, 1977; Klinger, Rosenfeld, Isom, & Deckard, 2016; Loughlin & Flora, 2017; Ma, Graves, & Alvarado, 2019; Mentch, 2020; Ross, Winterhalder, & McElreath, 2021; Shjarback & Nix, 2020; Tregle et al., 2019; Wheeler, Phillips, Worrall, & Bishopp, 2017; Worrall, Bishopp, Zinser, Wheeler, & Phillips, 2018).

4.1.2 Shooter bias: The missing forces flaw

With respect to the second flaw, it is clear that experimental social psychologists have ignored the contexts of actual deadly force decisions and the multiple influences on group disparities in fatal shootings, including the behavior of citizens themselves and whether such behavior varies across groups. There have been almost no serious attempts to connect experimental research to systematic analyses of fatal police shootings from the Criminal Justice literature, with nothing more than superficial citations of such research and no substantive input on how studies are designed or how research is conducted. Indeed, nearly a decade passed from the first publication using the FPST before researchers thought to ask about the very basic variable of neighborhood dangerousness (Correll et al., 2011), and 15 years passed before experimental social psychologists asked about whether different violent crime rates play a role in explaining racial disparities (Cesario et al., 2019; Scott, Ma, Sadler, & Correll, 2017).

In shooter bias studies, Black and White targets are shown holding guns with the same frequency; in other words, they are presented in equal proportions in those situations for which deadly force is relevant. The logic is that, if experimental participants are more likely to shoot Black targets in the FPST, then this same racial bias in the heads of police officers explains the per capita racial disparity in being shot. Yet for the results to apply, it must be the case that Black and White citizens are present in deadly force situations with equal likelihoods in the real world, otherwise factors such as differential exposure to the police may be sufficient to explain racial disparities.

In contrast to the underlying assumption in experimental studies, there is clear evidence that (1) the context of violent crime is an overwhelming influence on officers' decisions to shoot and (2) violent crime rates differ across racial groups (e.g., Barnes, Jorgensen, Beaver, Boutwell, & Wright, 2015; Cesario et al., 2019; Klinger et al., 2016; Ma et al., 2019; Miller et al., 2017; Nix, Campbell, Byers, & Alpert, 2017; Tregle et al., 2019; Wheeler et al., 2017; Worrall et al., 2018). Police officers do not use deadly force equally across all policing situations. The modal police shooting is one in which officers have been called by dispatch to the scene of a possible crime and are confronted with an armed citizen posing a deadly threat to the officer or to other citizens. It is also the case that violent crime rates differ very starkly across racial groups. Indeed, recent study suggests that the different rates of exposure to police through violent crime situations greatly – if not entirely – accounts for the overall *per capita* disparities in being fatally shot by police (Cesario et al., 2019; Fryer, 2016; Mentch, 2020; Ross et al., 2021; Tregle et al., 2019).

Once fatal police shootings are understood from this angle, it becomes clear that social psychologists have misunderstood this topic in their experimental approaches. Rather than first studying the nature of police shootings and then building experimental investigations around that understanding, researchers instead first created experimental worlds in which all group members are equal, under the assumption that this matched the actual behavior of groups and that their experimental findings would shed light on the disparate outcomes of those group members.

When it comes to explaining group disparities, researchers clearly prioritize their experimental findings over other possible causal forces on group outcomes. For example, of 18 recently published papers on shooter bias from experimental social psychology, only two raise the possibility that different behaviors of Black and White citizens might play a role in Black citizens' overrepresentation in being shot by the police (a possibility dismissed in one paper with indirect evidence and dismissed in the other paper with reference to a single article). This was true even when authors recognized that behavioral differences might account for other disparities, such as how the greater aggressiveness and criminality of men account for why they are more likely to be shot than women (Plant, Goplen, & Kunstman, 2011).

An important point concerning "blaming the victim" needs to be raised here, and this applies not only to fatal shootings but to all disparities. It is necessary to keep causal analysis distinct from "blaming the victim," or in Felson's (1991) terms, to not use a

https://doi.org/10.1017/S0140525X21000017 Published online by Cambridge University Press

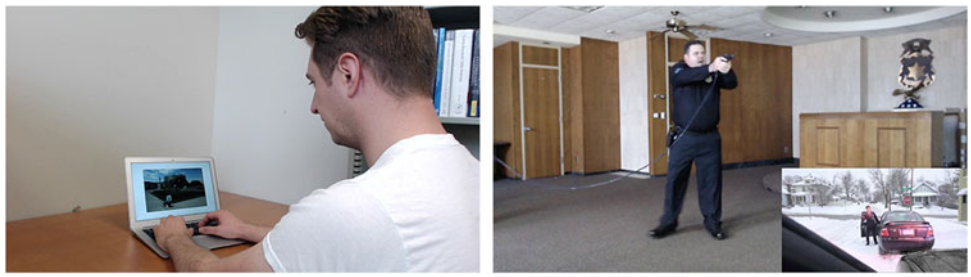

Figure 3. Left panel: Participant completing the standard laboratory First-Person Shooter Task. Right panel: A participant-officer completing an immersive shooting simulator, with video from officer's perspective superimposed in lower right corner.

blame analysis framework where a causal analysis framework is needed. Whatever the causal factors that lead an individual to one or another outcome, such factors can be described without the language of blame and responsibility. To say that a proximate cause of police shootings is involvement in crime is not to cast blame on a person for their own shooting, and certainly such an explanation should not be misapplied to those cases where criminal involvement is not present. But neither should a person's behavior be off-limits as part of a causal analysis merely because that person belongs to a minority group.

4.1.3 Shooter bias: The missing contingencies flaw

Research on shooter bias clearly illustrates the third flaw, the lack of attention to experimental contingencies and whether there are differences in motivation and ability between experimental and real-world decision-makers. Evidence of racial bias is reliably obtained with untrained citizens completing the FPST (Mekawi & Bresin, 2015), but the task has specific parameters that are required for such bias to be realized. For example, in the FPST, the target appears on the screen holding an object and the participant must make a decision within a response window relative to target onset. Thus, target race and object are presented simultaneously and responses after, say, 650 ms are considered errors.

The important question is whether these contingencies match the nature of actual police shootings. They do not. Officers almost always have information about citizen race much, much sooner than when the decision to shoot is made (and certainly well outside the window for ruling out controlled processing), and officers almost always have some interaction with the citizen before deciding to shoot. As noted above, experimental FPST participants are also given zero information about the situation surrounding the decision, a fact that matches no police shooting.

More important, the FPST is a task about *misidentifying harmless objects for weapons*. However, evidence of racial bias in the FPST has been used to make claims about widespread police officer bias in the decision to shoot. What has not been questioned is the degree to which fatal police shootings are actually *about* misidentification of harmless objects. If police shootings rarely involve the misidentification of objects under neutral conditions (which is the focus of the FPST), then it might be misleading to apply findings from the FPST to explain racial bias in fatal shootings more broadly. In fact, we estimated that the number of fatal shootings in which officers misidentify harmless objects for weapons is around 30 incidents per year (Cesario et al., 2019). To the extent that error rates on the FPST are informative for understanding racial bias, the task may be applicable only to an extremely infrequent event within a much larger set of related events. Indeed, considering that there are over 75,000,000 police–citizen contacts per year (Davis, Whyde, & Langton, 2018), this suggests the error rate for officers misidentifying a harmless object as a weapon – *the* central question of the FPST – is on the order of less than one in a million.

One could salvage the FPST by replying that the task still tells us something important about officers' decisions during these very infrequent events. Moreover, infrequent events can be tremendously important, and the tragic cases where an officer makes a clear error and shoots a citizen reaching for his wallet *are* the events that we as citizens should care the most about. However, two problems remain. First, the most reliable effect in the FPST is on response times and not on error rates; meta-analysis indicates that there is not a reliable effect of target race on shooting unarmed targets (Mekawi & Bresin, 2015). Second, such an argument requires ignoring the problems described above, which can change the applicability of such results to real-world cases. For example, the FPST assumes equal encounter rates with the police (as 50% of trials are White targets and 50% of trials are Black targets). However, if officers have differential contact with Black citizens (because of bias in discretionary stopping of citizens or simply because of different violation rates between Black and White citizens), then racial disparity in being shot while reaching for a wallet may exist while officers show no bias in the actual decision to shoot. A constant, race-blind error rate on the part of the police would still result in a greater proportion of Black Americans being shot while reaching for their wallets (see Cesario, 2021; Ross et al., 2021).

What about the failure to consider possible motivation and ability differences between real-world and experimental decision-makers? Social psychologists have overwhelmingly used convenience samples of naive undergraduates to study the decision to shoot (see Cesario & Carrillo, in press), participants for whom the decision is inconsequential and who have no training in how to make such a decision. Yet police officers typically receive over 1,000 hours of use of force training (Morrison, 2006; Stickle, 2016). It would be surprising if the ability to detect and classify objects, and the cognitive processes underlying such performance, was similar for experienced officers and undergraduates who have never made a single such decision in their lives. Interestingly, Correll et al. (2002) issued exactly this caution in the very first study on the FPST ("it is not yet clear that Shooter Bias actually exists among police officers … there is no reason to assume that this effect will generalize beyond [lay samples]," p. 1328). Yet despite this and later warnings (Cox & Devine, 2016), researchers continued to apply studies from undergraduates to police officers, even as data came to light that police officers did not show the same bias (e.g., Correll et al., 2007, 2014).

https://doi.org/10.1017/S0140525X21000017 Published online by Cambridge University Press

The fact that trained officers may use information in the decision landscape differently than untrained undergraduates represents an important ability difference between the two groups. If experts attend to different decision components or use these components differently than novices, *and this difference changes the effect of target race on the ultimate decision*, then the conclusion of widespread race bias in officers' deadly force decisions based on findings from undergraduate participants will be unwarranted. Indeed, sworn officers typically show little to no bias in the behavioral decision to shoot with the standard FPST (e.g., Akinola, 2009; Correll et al., 2007; Johnson et al., 2018; Ma & Correll, 2011; Sim, Correll, & Sadler, 2013; Taylor, 2011), and this is especially true for studies using immersive shooting simulators such as the one described above (e.g., James et al., 2013, 2014, 2016). Cesario and Carrillo (in press) summarized studies in which sworn officers completed the standard FPST and found that out of 64 possible tests for racial bias, only ~25% showed anti-Black bias whereas ~70% showed *no* bias on the part of officers in one direction or the other.

As a direct means of demonstrating the importance of collecting data with trained experts rather than naive undergraduates, Johnson et al. (2018) tested for differences between officers and students in the underlying cognitive dynamics of the decision to shoot. Was there evidence that trained versus untrained individuals were making the decision in a different way or using race differently during the decision process? Trained officers and untrained undergraduates completed the standard laboratory FPST. These researchers then modeled the data from each group using a drift diffusion model. This model describes the decision to shoot as a sequential sampling process in which people start with a prior bias to shoot or not and accumulate evidence over time until a threshold required for a decision is reached. The details regarding this modeling can be found elsewhere (see Johnson, Hopwood, Cesario, & Pleskac, 2017; Pleskac et al., 2018); for now, the important point is that the model allows for an understanding of how the cognitive processes underlying to the decision to shoot might vary between untrained and trained participants.

In these data, trained officers showed no racial bias in their behavioral decisions, despite untrained undergraduates showing such bias. More important, cognitive modeling of the decision data revealed *why* officers did not show bias in their behavioral responses. Officers showed two major differences compared to untrained undergraduates in the underlying decision components. First, race did not affect the manner in which officers accumulated evidence about whether to shoot. For untrained undergraduates, their processing of the object held by the target was "contaminated" by the target's race: When a harmless object was held by a Black target, the processing of his race interfered with processing of the object being held, pushing participants toward a "shoot" decision (resulting in more false alarms). Officers showed no such effect of race. They were able to extract information about the object in the person's hand independent of the target's race. Second, officers set higher thresholds for making a decision, accumulating more evidence before making a decision. In combination, these two components eliminated the effect of race on officers' behavioral decisions, an effect robustly observed in untrained participants.

Among trained officers, then, the decision process operated differently and race did not have the same effects on the underlying decision components as it did on untrained participants. Failure to understand or appreciate these differences leads researchers to inappropriately apply the results from undergraduates – who have no training and have never had to make such a decision before entering the lab – to expert decision-makers.

4.2 Implicit bias and group disparities

It would be difficult to find a concept from experimental social psychology that has spread more quickly and widely outside academia than implicit bias. There is no question that implicit bias research (1) has been used to explain why groups in contemporary American society obtain unequal outcomes and (2) has relied almost exclusively on studies using indirect measures such as the Implicit Association Task (IAT).[3] Other writings have critiqued the theoretical and measurement aspects of implicit bias research (Arkes & Tetlock, 2004; Blanton & Jaccard, 2008; Blanton, Jaccard, Gonzales, & Christie, 2006; Blanton et al., 2009; Blanton, Jaccard, Strauts, Mitchell, & Tetlock, 2015; Corneille & Hütter, 2020; Fiedler, Messner, & Bluemke, 2006; Mitchell, 2018; Oswald, Mitchell, Blanton, Jaccard, & Tetlock, 2013; Schimmack, 2020), so I restrict my discussion of this topic to those aspects most relevant to the question of explaining group disparities.

4.2.1 Implicit bias: The missing information flaw

In prototypical implicit bias research, as in studies using the IAT or other indirect measurement techniques (Fazio, Jackson, Dunton, & Williams, 1995; Greenwald, McGhee, & Schwartz, 1998), every possible source of information which could impact a person's judgment and behavior is stripped from the measurement of these unconscious or uncontrollable processes. In the best-case scenario, participants are shown cropped photos of faces belonging to different group members and make rapid categorizations of these faces; in the worst-case scenario, there are no group members whatsoever and group labels (e.g., "Black") serve as target stimuli instead. No information other than category membership is available to participants and button-press differences on the order of a fraction of a second are the outcome of interest. Additionally, research has shown that implicit or indirect measures can be sensitive to context information (see, e.g., Barden, Maddux, Petty, & Brewer, 2004; Blair, 2002; Gawronski, 2019; Gawronski & Sritharan, 2010; Wittenbrink, Judd, & Park, 2001). The fact that humans exist and are perceived only in contexts, and not isolated against empty backgrounds, should prompt meaningful discussion about the degree to which such context-less implicit bias measures will predict bias in real decisions.

4.2.2 Implicit bias: The missing forces flaw

Implicit bias research also reflects the second flaw outlined in this paper, which is that the effects of implicit bias are not appropriately compared to other influences on group outcomes. Consider the example of sex differences in STEM participation. Women and men do not have identical profiles of ability and interest relevant to STEM performance, and much research has explored the implications of these factors (Benbow & Stanley, 1980; Benbow et al., 2000; Ceci & Williams, 2010; 2011; Ceci, Williams, & Barnett, 2009; Cheng, 2020; Cortés & Pan, 2020; Hakim, 2006; Halpern et al., 2007; Kleven, Landais, & Søgaard, 2019; Kleven, Landais, Posch, Steinhauer, & Zweimüller, 2020; Lubinski & Benbow, 1992; Su & Rounds, 2015; Su et al., 2009; Valla & Ceci, 2014). How it is that millisecond differences in measured associations correspond to those factors which impact group disparities is questionable, given the lack of integration of such differences into the larger dynamics of STEM engagement and

https://doi.org/10.1017/S0140525X21000017 Published online by Cambridge University Press

performance. Although there is variation in the published literature, studies claiming to demonstrate the importance of implicit bias in explaining group outcomes often do not measure these other forces at all (e.g., Cvencek, Greenwald, & Meltzoff, 2011a; Cvencek, Meltzoff, & Greenwald, 2011b), do not compare the size of implicit bias effects to the size of these other forces (e.g., Nosek & Smyth, 2011), treat these other forces as control variables without directly comparing the size of implicit bias effects to these variables (e.g., Kiefer & Sekaquaptewa, 2007), or treat such forces as a predicted variable resulting from implicit bias rather than the reverse (e.g., Nosek, Banaji, & Greenwald, 2002; Nosek et al., 2009).

4.2.3 Implicit bias: The missing contingencies flaw

As for the third flaw, there has been a striking failure to explore whether the precise experimental contingencies required to demonstrate implicit bias in the lab correspond in some reasonable way to the contingencies present during real-life decisions. These contingencies include the twin features of the lack of ability and motivation, as well as the specific experimental details needed to reveal bias on indirect measures.

Consider some of the necessary experimental contingencies required both for the measurement of implicit cognition and for observing the effects of implicit bias on decision-making and behavior. Perhaps *the* central defining feature of implicit cognition is awareness (Greenwald & Banaji, 1995), and as such implicit measures are supposed to "neither inform the subject of what is being assessed nor request self-report concerning it" (p. 5).[4] A first-order, foundational question then is whether people are aware of their biases, aware of what is being assessed during the measurement of these biases, or aware of the effects of their biases. After all, if one defines implicit bias as discrimination based on "unconscious" processes and argues that implicit bias is so important as to have implications for legal doctrine in the United States (Greenwald & Krieger, 2006; Kang & Banaji, 2006), then certainly the basic question of awareness must have been thoroughly settled by now. As Gawronski (2019) describes, however, there is currently *no convincing evidence* that people are uniquely unaware of their biases or the effects of their biases.[5] It is striking that the concept of implicit bias has been pushed into federal policy at the highest levels of the U.S. government without any convincing evidence concerning even basic questions about the measurement or the effects of implicit bias.

Indirect measurement techniques (as a means of assessing stereotype associations) require specific contingencies to reveal bias on the part of participants. Take for example the IAT (Greenwald et al., 1998). As with other control tasks, such as the Stroop task, no one shows bias in their decisions if given sufficient time to respond.[6] Thus, a speeded response is a *required condition* for measurement of implicit bias so that controlled cognitive processes will be prevented or attenuated from impacting responses. In this way, implicit measures can assess "implicit attitudes by measuring their underlying automatic evaluation" (Greenwald et al., 1998, p. 1464), as opposed to measuring a more controlled evaluation elicited by the stimulus.

In addition to measurement, there are also necessary conditions to demonstrate the effects of implicit bias on behavior and decision-making. Consider the central claim by implicit bias researchers that automatically activated associations influence us even when we don't want them to (e.g., "implicit biases are especially intriguing, and also especially problematic, because they can produce behavior that diverges from a person's avowed or endorsed beliefs or principles," Greenwald & Krieger, 2006, p. 951). Given this, people must be in a decision situation where controlled processes – that is, *what we want* – cannot play a role, conditions where we want to respond in unbiased ways but are unable to do so. This requires that a person lacks the ability to exercise controlled processes, as in a decision with a short response time window. Without this feature, the decision situation is no longer one in which we are unable to produce the desired, unbiased response. Good experimental practice and inference would then require that, in implicit bias research, both contingencies are in place: People do not *want* to be influenced by categorical information but are in decision situations where such controlled processes cannot influence responses. Given that none of the studies recently presented as strong evidence for the behavioral prediction of implicit bias ensured that these contingencies were met suggests that this practice is not widespread (Jost, 2019).

As a final, critical contingency, as noted earlier there is overwhelming evidence that categorical bias is overridden when decision-makers are provided with individuating information (e.g., Kunda & Thagard, 1996). In the measurement of implicit bias, no individuating information is ever presented; yet it is common to apply laboratory findings of implicit bias to real decisions which contain strong individuating information, such as in hiring decisions or decisions about one's own career choice (where a person clearly has interest and ability information; e.g., Nosek & Smyth, 2011).

In terms of explaining group disparities, it follows that the bulk of the underrepresentation for any group must be because of an underrepresentation of people who are *ambiguous* with respect to their performance at the task at hand, because it is only these people for whom decisions will be affected by implicit bias on the part of decision-makers. In the case of "gatekeepers" making biased decisions against potential STEM students (Nosek & Smyth, 2011), the "A" student and the "F" student are both unaffected by implicit bias on the part of the guidance counselor (because there is unambiguous positive and negative individuating information, respectively). This means that the sex disparity must be comprised of "C" students who *would* have become successful in STEM careers had implicit bias not caused the guidance counselor to unintentionally steer those students out of a STEM track. It is the responsibility of implicit bias proponents to show this is the case.

Implicit bias research, then, provides another example of the fundamental weaknesses of experimental social psychology when explaining group disparities. Without providing any relevant information to participants, researchers obtain evidence of the biasing effects of category information. Such associations as measured by millisecond response time differences – obtained under completely discordant conditions to the real world and which do not correspond to the presumed psychological constructs of interest in a straightforward way (see, e.g., Blanton, Jaccard, Christie, & Gonzales, 2007; Uhlmann, Brescoll, & Paluck, 2006) – are proposed to explain complex and sizable group disparities. Little effort is made to integrate these differences into a detailed model which includes other, strong influences on outcomes or specification of the real-world performance parameters. These weaknesses are consistent with the poor performance of implicit bias measures to predict discriminatory behavior (see, e.g., Blanton et al., 2009; Oswald et al., 2013, 2015).[7]

https://doi.org/10.1017/S0140525X21000017 Published online by Cambridge University Press

4.3 Racial disparities in school disciplinary outcomes

A final example is the recent study in experimental social psychology on racial disparities in school disciplinary outcomes. There are well-known racial disparities in suspensions and expulsions, with Black schoolchildren more likely to receive such outcomes than White, Hispanic, or Asian schoolchildren (Lhamon & Samuels, 2014). At issue is why this per capita disparity exists and whether distorted interpretation of behavior because of racial stereotypes explains such disparities. That is, are schoolteachers interpreting the same behavior on the part of Black and White schoolchildren differently and, therefore, referring them for disciplinary action at different rates, even while the behavior of Black and White kids is the same?

Experimental social psychologists have followed the familiar pattern of instructing participants to make punitive judgments of hypothetical schoolchildren from simple written scenarios, with targets who are presented as equal on every dimension other than their race. After observing an effect of target race on disciplinary decisions, researchers then loop back and claim that such findings can help explain why racial disparities in real classrooms exist (Jarvis & Okonofua, 2020; Okonofua & Eberhardt, 2015).

An analysis of this research reveals the three flaws identified above. The information provided to participants in these experimental studies are impoverished descriptions of real teacher–child experiences, removing important information that real decision-makers could use, such as a child's history of behavior in the classroom, the other children involved, the teacher's current intentions and behavior, or even the general context surrounding the event. All the knowledge that the teacher has concerning the student's history and past behavior simply cannot play a role in their experimental judgments. This is important because the distribution of student disciplinary action is highly skewed and is principally tied to specific students; the question is not about the average, generic student but about specific students at the tail end of a distribution. For example, in one large survey of teacher referrals for disciplinary action, 93% of the 22,000 students recorded did not receive a single referral, 4% received only one referral, and *six students* received more than 20 referrals each (Rocque & Paternoster, 2011). Experiments are about group average effects (e.g., "Does a sample of participants show an average difference in disciplining unknown, nonspecific Black or White students?"), but the distribution of disciplinary actions suggest this misses the nature of the actual topic under study.

Researchers prevent teachers from making unbiased judgments if such information plays a strong role in real decisions and forces participants to use the only diagnostic information given to them. For example, studies on race and classroom discipline give teachers a student's name (manipulated to be either a common Black or White name) and a one-paragraph description of an event ("You tell DeShawn to pick his head up and get to work. He only picks his head up"). Whether these vignettes contain information used by teachers when making real disciplinary decisions is unknown.

These experimental designs also fail to consider the possible influence of other factors that may play a role in a child's behavior, such as socioeconomic status, family structure, cultural norms for the teacher–child relationship, parental expectations, interest in school, delay of gratification, and so on, all of which differ across racial groups and would reasonably be expected to relate to behavioral differences in the classroom (Andreoni et al., 2019; DeNavas-Walt, Proctor, & Smith, 2013; Heriot & Somin, 2018; Hsin & Xie, 2014; McLanahan & Percheski, 2008; Musu-Gillette et al., 2018; Price-Williams & Ramirez, 1974; Rocque & Paternoster, 2011; Wright et al., 2014; Zytkoskee et al., 1971). Whatever the size of participants' racial bias in disciplining hypothetical Black versus White schoolchildren in an experimental situation, one cannot draw any conclusions about whether such categorical biases impact disciplinary outcomes in the real world because the experimental bias effect is not understood in relation to these other factors. An assumption justifying the design of such studies is the expectation that children who differ in myriad important ways should behave identically in the classroom.

As support for this claim, consider a recent paper on race and school suspensions by experimental social psychologists, which begins by stating that racial differences in school suspension are "not fully explained by racial differences in socioeconomic status or in student misbehavior" (Okonofua & Eberhardt, 2015, p. 617). No report is given of how much the racial disparities *are* explained by these factors, just that some non-zero amount remains. As evidence for this claim, six citations are provided, but none of these citations measure student behavior and show that Black and White students are behaving similarly. Indeed, one of these citations states "The ideal test … would be to compare observed student behavior with school disciplinary data. Those data were not available for this study, nor are we aware of any other investigation that has directly observed student behaviors" (Skiba, Michael, Nardo, & Peterson, 2002, p. 325).[8] In contrast, Wright et al. (2014) did find that racial differences in school suspension rates were fully accounted for by prior behavioral problems of the student. The point is not to single out these researchers (as such claims are broadly made by nearly everyone doing similar research), but instead to illustrate an additional example of the problems identified above.

Moreover, using experimental social psychology to explain school suspensions and expulsions reflects the third flaw as well: A lack of attention to the actual contingencies needed to produce stereotyping effects in the lab and whether such contingencies resemble real-world situations. As noted earlier, stereotyping effects occur under conditions of ambiguity and are absent or small when perceivers have individuating information or are judging unambiguous behaviors. To the extent that teachers are misconstruing or misinterpreting students' behaviors because of stereotypes held about different racial groups, those effects are therefore predicted to occur in the absence of individuating information or for ambiguous behaviors. How categorical bias might reveal itself in long-term interactions such as teacher–student relationships, where plenty of individuating information is available, is not established.

Some study by the leading scholars within social psychology on school disciplinary disparities has tried to take a more dynamic perspective. For example, Okonofua, Walton, and Eberhardt (2016) propose that the teacher–student relationship can devolve over time and that initial stereotype effects can increase in strength as teachers' expectations and worry about minority students' behavior affects students' behavior in the classroom (see also Madon et al., 2018; Martell, Lane, & Emrich, 1996). Of course, whether initial teacher concerns about classroom management eventually lead Black students to enact those behaviors that would get them expelled, when they would not have otherwise done so absent such expectations, is unclear. Nor are the effects of such expectations set within the context

https://doi.org/10.1017/S0140525X21000017 Published online by Cambridge University Press

and force of the other strong effects listed earlier on students' outcomes.

5. What *do* experimental studies of bias tell us?

To say that studies in experimental social psychology cannot tell us about real-world group disparities is not to say that such studies are worthless. These studies provide a wealth of information about the function and process of storing and using categorical information. However, if researchers want to know about real-world group disparities, such findings cannot provide them with the information they seek.

The standard way of interpreting experimental stereotyping findings has already been described: Experimental evidence that participants are biased against identical targets from different groups reflects the power of stereotypes to affect individual decision-makers. The assumption that the same processes operate in the real world means that removing decision-maker bias will result in groups obtaining roughly similar (or at least substantially more similar) outcomes.

Yet is this interpretation the correct one? An alternative interpretation of the results of experimental studies of bias starts with the understanding that people learn the conditional probabilities of the behavior of different groups as they navigate their social worlds. In other words, groups differ in their characteristics and people pick up on this, storing diagnostic information about relative group differences even if imperfectly so (Eagly, Wood, & Diekman, 2000; Eagly, Nater, Miller, Kaufmann, & Sczesny, 2020; Jussim et al., 2009, 2015a, 2015c; Koenig & Eagly, 2014; McCauley, Stitt, & Segal, 1980).

Then, they enter a social psychology experiment on bias. They are asked to render a judgment about a target without being given diagnostic or distinguishing individuating information. Under such conditions, they end up using the information that they have come to learn as being probabilistically accurate in their daily lives, and categorical influence dominates.

Thus, through a kind of *methodological trickery*, the experimenter has created a world in which information that is probabilistically predictive in everyday life becomes completely inaccurate given the systematic design of our experiments. This interpretation is consistent with a view of stereotyping that describe perceivers as forming conditional probabilities and emphasizes how categorical effects are most likely under conditions of ambiguity and uncertainty, when no strong individuating information is present (Krueger & Rothbart, 1988; Kunda & Thagard, 1996; Lick, Alter, & Freeman, 2018; McCauley et al., 1980). Given the design of most experiments, it is not surprising that there are decades of laboratory studies showing stereotyping effects. To be clear, this provides no information about whether this type of categorical influence leads to disparate outcomes across groups. It does reveal that experimenters are skilled at creating worlds whose landscapes do not match the real world in any way, and participants fail to behave perfectly according to the standards of the experimenter when placed in such worlds.

In light of this reframing, what does the standard interpretation of experimental studies reveal about researchers' assumptions of how minds should and do operate? Throughout this paper, I have noted that the standard experimental design presents targets "who vary only with respect to the social categories to which they belong." What do researchers intend when they design stimuli in this way? In doing this, researchers intend to make targets equal *on all dimensions relevant to the decision at hand*. For example, in the FPST, the single relevant piece of information in the decision to "shoot" is whether the target is holding a gun or not. If participants are influenced by anything other than the object in the target's hand, then researchers conclude that participants are making erroneous decisions – that is, they are showing bias. This includes cases when participants are influenced by factors related to a person's race that are probabilistically related to threat or handgun use, for example, having been previously arrested for a violent crime. Similarly, in studies of STEM hiring, the single relevant piece of information is the qualification of the applicant as revealed by the resume; being influenced by anything other than this information is treated as biased, erroneous decision-making.

What this illustrates is the researcher's belief that *participants are wrong to use any information other than the information deemed relevant by the researcher*. This includes information that the participant has learned prior to entering the experiment, information that may be probabilistically accurate in everyday life. In the mind of the researcher, participants should not use information within the experiment that may actually lead to *more accurate* decisions outside the experiment – not because such information is reliably incorrect, but because the experimenter has artificially made it incorrect. The researcher demands that participants are accurate as defined by the decision landscape of the experiment, no matter how disconnected this landscape is from the real world. Researchers, thus, require a kind of *blank slate worldism* of their participants in judging accuracy and bias, where information from one world must be erased when moving to the next. Such a demand on the part of social psychologists in fact violates a core tenet of good prediction, which is the use of priors in updating posterior prediction. Bayes' rule would require participants in social psychology experiments to include the target's categorical information in their judgments (though of course the effect of categorical information should depend on the strength of the data, as it does).

6. Broader consequences

Beyond the specific conclusions about group disparities, experimental social psychology has had a significant – and potentially misleading – impact on broader questions about the human mind and human nature. This research has led directly to the widespread attention currently given to the topic of implicit bias. Originally, dual process models in social psychology supported a satisficer view of the human mind, one in which people did "good enough" (and were thus subject to bias) unless motivation and ability were high (Fazio, 1990; Fiske & Neuberg, 1990; Petty & Wegener, 1999; Smith & DeCoster, 2000). Importantly, such models were explicit that biasing effects were conditional (Bargh, 1989); they were not present at all times and for all people.

As experimental studies of categorical bias proliferated and as demonstrations of bias became more attractive than demonstrations of accuracy (e.g., Higgins & Bargh, 1987; Jussim, 2012b; Jussim et al., 2009), the published literature left one with the impression of widespread, inescapable error in decision-making and the important point that bias occurs *under specific experimental conditions* was given a backseat to the more attractive story of widespread bias in real-world decisions (Greenwald & Krieger, 2006). Moreover, as social psychology moved further away from actual behavior and increasingly focused on millisecond reaction times, whether such differences mattered for actual decisions became increasingly unclear.

https://doi.org/10.1017/S0140525X21000017 Published online by Cambridge University Press

At the same time, demographic groups in the United States continued to obtain unequal outcomes despite little overt, official discrimination for several decades (and in places such as academia, preferential policies in favor of underrepresented groups), coupled with increasingly egalitarian attitudes. These disparities presented a puzzle. If groups were not being overtly barred from entry and decision-makers widely expressed egalitarian beliefs, what was causing persistent disparities?

Enter the concept of implicit bias, supported by experimental social psychology studies on categorical bias (Greenwald & Banaji, 1995). As this research was taken up by people outside the research community, the understanding of the human mind morphed from "under certain conditions, bias may emerge" to "unconscious bias is ever-present and impossible to control," with a lack of attention to those studies showing individual variation in automatically activated concepts (e.g., Fazio et al., 1995). By now, this view is ubiquitous and claims of uncontrollable, unavoidable, pervasive, unconscious bias can be found anywhere one cares to look.

Such a view of the human mind, however, is in no way justified by the experimental studies on which it is built. There is so little overlap between our experimental parameters and the parameters of real-world decisions that the popular view of the human mind as swamped with uncontrollable bias is premature. It is troubling that researchers have not devoted serious research attention to exploring this gap.

At the same time that social psychologists have been using their findings to explain group disparities, people outside academia have enthusiastically adopted these claims. This has been true throughout popular culture, government organizations, the legal system, and the corporate world. In the case of police shootings, the claim that implicit bias is responsible for racial disparities is widely broadcast in newspaper accounts of fatal police shootings, with studies from experimental social psychological cited as evidence (e.g., Carey & Goode, 2016; Dreifus, 2015; Kristof, 2014; Lopez, 2017). In the case of school disciplinary disparities, President Obama's 2014 "Dear Colleague" letter on the "Nondiscriminatory Administration of School Discipline" was explicit in rejecting the idea that actual behavioral differences across racial groups contribute meaningfully to the corresponding disparities in school suspensions. It also named implicit bias training as a possible solution for ensuring that school police administer discipline in a non-discriminatory manner. It is difficult to overstate how widespread this belief has become in the last decade, driven primarily if not wholly by research from experimental social psychologists. Indeed, some researchers have actively pushed this agenda, appearing on televised news programs, holding press conferences, writing advocacy pieces, and testifying in court (as described in, e.g., Mitchell, 2018).

7. Related critiques

Although I focus on social psychology experiments in this paper, related critiques have been made in other literatures. A brief review of these critiques, some of which are general methodological critiques and some of which are specific to group disparities, provides additional support to the current argument.

On the question of group disparities specifically, Heckman's (1998) analysis of racial and gender disparities in employment supports the current analysis. In typical "audit studies" (e.g., Bertrand & Mullainathan, 2004), a set of prospective employers are sent resumes that are identical except for the race of the applicant; research typically finds that Black applicants receive fewer callbacks for interviews than White applicants. Such findings are then used as evidence that actual racial disparities in employment are because of discrimination on the part of employers. Thus, the general format of experimental labor market studies is the same as the social psychology research described in the current paper: If we can show average levels of race-based differential treatment between hypothetical people who are otherwise presented as equal, then this same differential treatment is responsible for actual group disparities.

Heckman argued that average levels of market-wide discrimination cannot necessarily be applied to real people engaged in real transactions, because such transactions do not occur at the market-wide level. Employment transactions are between specific people and specific firms, and *if the people and firms in experimental studies do not match the characteristics of real people and firms in the market*, then experimental results are irrelevant for explaining real group disparities. Suppose an experimental audit study finds that employers at Goldman Sachs engage in discrimination against Black applicants. If it is the case that Black applicants do not apply to Goldman Sachs, or that actual Black applicants do not have the resumes that would make them competitive at Goldman Sachs, then whether employers at Goldman Sachs discriminate against artificial Black applicants tells us nothing about why Blacks may be under-employed there or anywhere else in the financial market.

There is the same problem in labor market studies as in studies in experimental social psychology: A lack of attention to the degree of overlap between the characteristics of real group members and the characteristics of our hypothetical experimental targets. And this failure, as in social psychology, distorts our understanding of the nature of group disparities. As Heckman summarized, "A careful reading of the entire body of available evidence confirms that most of the disparity in earnings between blacks and whites in the labor market of the 1990s is due to the differences in skills they bring to the market, and not to discrimination within the labor market" (p. 101; see also Neal & Johnson, 1996).

In terms of broad methodological critiques, similar concerns have been raised in the field of judgment and decision-making (JDM). Hogarth (1981), for example, highlighted the discrepancy between the discrete judgments used in experimental JDM research and the continuous, interactive judgments frequently found in the real world. He used this discrepancy to highlight how researchers' failure to incorporate the role of feedback in experimental decision tasks could lead to distorted conclusions. Specifically, he demonstrated that decisions characterized as "biased" in discrete judgments could be understood as functional when decisions were continuous. Similarly, a major thrust of Gigerenzer and colleagues' research program has been to show that the structure of the decision environment is a crucial consideration for a full understanding of accurate and inaccurate decisions. Failure to appreciate the relation between the organism and its environment can lead to misleading conclusions about the nature of human rationality and decision-making (Dhami, Hertwig, & Hoffrage, 2004; Gigerenzer, Hoffrage, & Kleinbölting, 1991; Pleskac & Hertwig, 2014). Tetlock (1985) also analyzed the nature of JDM research and noted how laboratory studies lacked *accountability* for decision-makers, a key component inherent to most real-world decisions and one which can change the nature of decisions. Thus, there is precedent for being concerned about social psychologists' lack of interest in the degree to which their experimental tasks reflect the decision landscape in which actual decisions are made or whether the characteristics of real decision-makers match those in our experimental settings.

https://doi.org/10.1017/S0140525X21000017 Published online by Cambridge University Press

Relatedly, Eagly and colleagues' study on gender differences and leadership style provide supportive evidence for the arguments advanced here (Eagly & Johannesen-Schmidt, 2001; Eagly & Johnson, 1990). These researchers found that some gender differences in leadership style were larger in laboratory studies compared to studies conducted in actual organizational settings. The explanation for this difference in methodology could be understood in the terms described here, which is the failure to include real-world information in our laboratory studies. Specifically, actual roles in organizational settings contain role requirements, which can exert powerful effects on behavior regardless of the person occupying the role. In laboratory studies, in contrast, this influence is absent, hence the greater potential for gender to exert an influence on leadership behavior in this context.

To be fair, within social psychology there are some lines of research on stereotyping and disparate outcomes that do consider group behavioral differences as an important part of the causal chain producing group disparities. For example, Diekman et al. (2017) have proposed a goal congruity model to help understand sex differences in STEM participation. In this model, the communal goals that people have, in combination with their beliefs about how different STEM and non-STEM careers can fulfill those needs, impact STEM engagement and ultimately career choice. Importantly, this model accounts for at least some of the sex disparities in STEM participation by taking seriously the sizable male–female difference in communal goals.

Finally, the current study is most closely related to broad concerns in the experimental literature on external validity. Part of the current analysis raises multiple concerns regarding the external validity of experimental social psychology, and this is certainly not new. However, this study goes beyond past treatments in several ways. First, this paper outlines which features of the typical experimental investigations are threats to external validity and analyzes how the fallacies and assumptions underlying researchers' approaches to the question of group disparities directly lead to choices that undermine external validity. Second, the current study is not a broad indictment of the external validity of typical experimental social psychology. The standard experimental social psychology study can tell us much about how categorical information is formed and used, and I raise no issue with the external validity of those studies. Instead, the concern here is specifically with the use of these findings to explain real-world disparate outcomes. Finally, the current study goes beyond typical external validity concerns because, even if the external validity of current studies was improved, the problems inherent to this approach are so fundamental that they still could not be applied to explain group disparities. For example, if distributional differences between men and women on STEM-related attributes are not taken into account when explaining group disparities in STEM participation, then irrespective of any changes to the experimental process researchers will still misunderstand the nature of this disparity. A way of thinking about the relationship between the current analysis and past critiques of external validity is that the current study uses those past critiques as a vehicle for a broader, more systematic dismantling of current experimental studies on bias.

On external validity, relevant data supporting the current argument come from Mitchell (2012), who compared effect sizes of laboratory studies to field studies. Although the relationship between the two was strong and positive, this varied by subfield in important ways. Social psychology not only had a lower correspondence between lab and field studies than some other subareas, but social psychology was also the subfield in which the *sign* of the effect reversed most often. Although the purpose of Mitchell's analysis was not to identify all the features that impact lab-field correlations, the relatively poor performance of social psychology could be understood with the current framework – to the extent that the lab studies fail along the three flaws outlined here, the correspondence of these experimental effects once behavior returns to the field will be low. Of course, not all the social psychology studies in Mitchell were of decision-maker bias, but other analyses have found similar, supportive effects (e.g., Eagly & Johnson, 1990; Koch et al., 2015).

8. A new (or at least rehashed) approach

If the current approach to understanding group disparities is not just misguided but fundamentally flawed, what might be an alternative, more productive research cycle? Although it would be nice to claim a completely new approach to studying these important topics, what follows is largely a rehashing and reemphasizing of other, better recommendations that have already been made, for example, by Dasgupta and Stout (2012) and Mortensen and Cialdini (2010), with some further elaboration and connection to other critiques from the past several decades. The major difference is that I begin by explicitly noting that in many (perhaps most) cases of studying group disparities, we may end up concluding that experimental social psychology cannot contribute or at least will play a distant backseat to other approaches.

Studies of group disparities on any outcome should begin first and foremost with a task analysis of the decision itself as it exists outside the laboratory. This would involve detailed discussions with those individuals responsible for making such decisions, ideally including novice and expert decision-makers. Researchers might also meaningfully enhance the quality of their models by completing training protocols themselves, to learn how the decision is supposed to unfold (at least as formally instructed). In the case of police shootings, beginning at this step would likely have led to a drastically different methodology used by experimental social psychologists, one which incorporated actual features of deadly force decisions.

The second step in the process involves the study of members of groups who are obtaining disparate outcome on the topic of interest (both more and less desirable outcomes), including behavioral, personality, or other individual differences relevant to the topic at hand. This can often be useful in confirming that the factors identified by decision-makers in step 1 are, in fact, relevant. This step is also important for placing any categorical bias effects in the context of the size of these performance-related differences. Beyond giving us a more accurate understanding of the nature of group disparities, this can also provide information about the strength of different interventions to reduce such disparities. The expectation about what the world will look like after eliminating all decision-maker bias is very different depending on whether there are no differences or large differences across groups.

In the case of shooter bias, an initial task analysis would have revealed that the context and behavior of the target citizen is critical and that the context of violent crime is a central part of the officer's decision to shoot. The second step would have led to the recognition that there are very sizable differences across groups in violent crime rates and led to an appreciation that any biasing effects of race on an officer's decision must be placed in the context of these behavioral differences. The same is true, for example, of intellectual performance differences across groups, where sometimes

https://doi.org/10.1017/S0140525X21000017 Published online by Cambridge University Press

average differences do not exist but differences are large at the extreme tails and other times average group differences do exist and are sizable (Ceci & Williams, 2010; Fryer & Levitt, 2010; Halpern et al., 2007; Hsin & Xie, 2014; Lubinski & Benbow, 1992). Outcome differences in demographic disparities among, for example, college grade point averages (GPAs), majors, and graduation rates must be understood in the context of these sizable incoming differences across racial and ethnic groups (e.g., ACT, 2017), and interventions that do not address these differences at the core are unlikely to stem the cascading and continuing differences over time.

Only after the first two non-experimental steps comes the third step of designing experiments informed by the data already obtained. This will almost always necessitate more involved and difficult studies with non-student samples; what follows would likely be a steep decline in both the number of studies conducted and the proportion of studies involving undergraduate convenience samples.

The final step in relating back to the real-world disparities of interest involves integrating the size of categorical effects from experimental tasks with the sizes of other effects on a group's outcomes, for example, behavioral and personality differences across groups. This is something that will be specific to the domain under study as it is unlikely that many of the same factors impact outcomes to the same extent across domains (but see Gottfredson, 1997, 1998, 2004).

This call for a new approach to research complements other, previous concerns about the approach of standard psychological science. Already noted are the proposals by Dasgupta and Stout (2012) and Mortensen and Cialdini (2010). Other recent examples include Rozin's (2009) assessment of how changes to the reward structure in psychology would improve the science. As he stated (emphasis added):

> In such cases, as with the nth study (where $n > 10$) on a particular phenomenon or claim, it is appropriate to determine whether proper controls have been conducted, whether alternative accounts have been dealt with, and whether there are any errors in thinking or experimentation. *But first, we have to find out what it is that we will be studying, what its properties are, and its generality outside of the laboratory and across cultures.*

Aligned with Rozin's critique, the current study pushes back against a movement that gained momentum with the emergence of social cognition in the late 1970s and perspectives such as Mook's "Defense of External Invalidity" (Mook, 1983). These forces pushed the importance of systematic design and justified the measurement of small differences in highly impoverished experimental settings, without consideration of whether the decisions made in these studies related in clear ways to the actual decisions that, ultimately, we care so much about (see also Ring, 1967). Another way of framing the problem is to suggest that social psychology has been more focused on publishing demonstrations of bias than on fully understanding the nature of group disparities through the pursuit of a "strong inference" model (Platt, 1964).

9. Conclusion

What can experimental social psychology tell us about why different segments of society are not evenly represented across all outcomes? Experimental studies of categorical bias can and do tell us about the functions and processes of storing group-based information. However, the disconnect between the experimental parameters of these studies and the conditions surrounding real-world decisions makes our experiments irrelevant when it comes to understanding the complex dynamics of group disparities. Of course, there is individual-level bias and discrimination; tribalism and intergroup bias are features of all human minds. But if the goal is to study systematic categorical bias and its effects on group outcomes, a different approach is needed. I describe one possible new approach for experimental social psychology, one which begins not with the assumptions of academic researchers holding the goal of demonstrating bias but instead with an analysis of the actual decision itself. Such an approach would not only change the relevance of social psychology for understanding group disparities, but may also correct some of the misleading claims about the human mind that have extended out from academia in the last two decades.

Acknowledgments. I thank Michael Bailey, E. Tory Higgins, Lee Jussim, Calvin Lai, Richard Lucas, and three anonymous colleagues for productive discussions and feedback on earlier drafts of this study. Alice Eagly and two anonymous reviewers provided outstanding and critical comments that greatly increased the quality of this manuscript. This study also benefitted from discussions with friends and colleagues at the Duck Conference on Social Cognition (2017).

Financial support. This study is based on work supported by the National Science Foundation under Grants No. 1230281 and 1756092.

Conflict of interest. None.

Notes

1. Although speculative, this claim is consistent with the expectation of many social psychologists that, absent biasing agents, all groups would attain roughly equal outcomes; that is, evidence of disparity is taken as evidence of discrimination (or at the very least it is taken as evidence that something is wrong and in need of fixing). Social psychologists continue to place experimental research in their narratives about why group disparities persist decades after explicit discrimination has been legally banned and attitudes have become markedly more egalitarian (e.g., Dovidio, 2001; Dovidio et al., 2008). Quantitative data confirm the strong ideological lopsidedness of academics, particularly in the social sciences (e.g., Haidt, 2011; Inbar & Lammers, 2012; Jussim, 2012a; Jussim, Crawford, Anglin, & Stevens, 2015b).
2. In response to a question about why different groups achieve different outcomes, Thomas Sowell reframed the question as: "I would look at it differently … I would say, 'Why would we expect different groups to do the same?' Americans have come here from all over the world, and why would you ever expect that countries that have entirely different histories, located in entirely different climates, different geographies … Why would you expect those countries to develop exactly the same mix of skills to exactly the same degree so that people would arrive on these shores in such a way that they would be represented evenly across the board? Nowhere in the world do you find this evenness that people use as a norm. And I find it fascinating that they will hold up as a norm something that has never been seen on this planet, and regard as an anomaly something that is seen in country after country."
3. A few quick examples from leading social psychologists illustrate these points:

- On racial disparities in criminal justice outcomes, Banks, Eberhardt, and Ross (2006) state: "The racial bias research centers on the Implicit Association Test (IAT), which aims to measure implicit bias that operates beyond individuals' conscious awareness, and may exist even among individuals who genuinely believe themselves to be unbiased" (p. 1170).
- On "the persistent disparity in economic, residential, and health status between Blacks and Whites," Dovidio et al. (2008) state: "less conscious and more indirect" … "racial biases … occur implicitly, without intention or awareness" and "are assessed with new techniques (e.g., response time measures) … which assess spontaneous and uncensored reactions" (pp. 478–479).
- On the role of trust "in the worlds of business, law, education, and medicine, and even more ordinary daily interactions between individuals," Stanley,

https://doi.org/10.1017/S0140525X21000017 Published online by Cambridge University Press

Sokol-Hessner, Banaji, and Phelps (2011) state that "implicitly held attitudes" (as measured by the IAT) have "very real cost for individuals and society" (pp. 7710, 7714).

4. The original definitions of implicit cognition within social psychology solely emphasized the lack of awareness rather than uncontrollability, as for example in Greenwald and Banaji's (1995) original definitions. However, definitions now emphasize both awareness and controllability, as in the definitions found on the Project Implicit website or, for example, in Nosek, Greenwald, and Banaji (2007): "The term implicit has come to be applied to measurement methods that avoid requiring introspective access, decrease the mental control available to produce the response, reduce the role of conscious intention, and reduce the role of self-reflective, deliberative processes" (p. 267).

5. On awareness of the contents themselves, Gawronski states: "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, people the available evidence suggests that people are aware of the mental contents" (p. 578). On awareness of the effects of those mental contents, Gawronski states: "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, the preliminary evidence that implicit, but not explicit, biases influence judgment outside of awareness is rather weak and prone to alternative interpretations" (p. 578). With respect to the IAT specifically, there is *nothing* in the measure that ensures that participants are unaware of their association or unaware of what is being assessed during the IAT. Indeed, the IAT is an attention-grabbing effect precisely because the person can consciously *feel* the difficulty of certain categorizations even while they do not want those categorizations to be more difficult.

6. This will be true for the behavioral response itself; that is, bias as measured by pressing one or another button or saying one or another color will be eliminated with unlimited response time windows. Bias as measured by response times may still be observed under longer response windows, but even this requires that participants do not have a *minimum* response time window restriction on their response (imposed by either themselves or the experimenter), for example, that one cannot respond before 2 s.

7. One of the strongest defenses of implicit bias to explain real-world group outcomes was mounted by Jost et al. (2009). These researchers listed 10 studies claiming to prove that "implicit bias is beyond reasonable doubt." Although a thorough evaluation of these studies is beyond the scope of the current paper, it is useful to address how these studies fare with respect to the three flaws described here. Despite the fact that the summary is explicitly designed to address implicit prejudice and group outcomes, only four of the cited 10 papers deal specifically with differential treatment of social groups based on implicit bias. Of these four, however, all exhibit at least one of the fatal flaws identified above and, therefore, none are convincing demonstrations of the role of implicit bias in explaining group disparities. Moreover, given that the stated purpose of such studies is to uncover biases of which people have "little or no awareness" (Jost et al., 2009, p. 40), the fact that *zero* of the 10 studies had any assessment of whether people were aware of their own associations or their effects further removes these papers from providing convincing evidence of the effects of implicit bias on group outcomes. Rooth (2007) used fully equated applications with the applicant name changed, failing to incorporate real group differences in estimating the size of categorical bias. Rudman and Glick (2001) had undergraduates with no training make hiring decisions. Plant and Peruche's (2005) research is a shooter bias study and has the problems described in section 4 of this paper. Green et al. (2007) did use actual physicians but rated hypothetical vignettes and in fact showed that *physicians with high anti-Black bias (as measured by the IAT) actually treated Black and White hypothetical patients equally*. von Hippel et al. (2008) do not relate to group disparities (and do not have any comparison groups to the target group under study, thus cannot provide evidence for group disparities). Arcuri, Castelli, Galdi, Zogmaister, and Amadori (2008), Palfai and Ostafin (2003), Rudman and Ashmore (2007), Gray, Brown, MacCulloch, Smith, and Snowden (2005), and Nock and Banaji (2007) all do not relate to group disparities. All but one study were almost certainly underpowered to detect the effects reported, given the sample sizes and designs.

8. To be fair, many of these studies do try other, indirect ways of establishing racial bias in teachers' interpretations of behavior, e.g., use of discriminant analyses. And it is important to be appropriately cautious about the uncertainty in whether there are behavioral differences in the types of behaviors performed by Black and White schoolchildren. Yet, at the same time that many authors have argued for a lack of evidence in behavioral differences, it is not always clear that the data support this claim. For example, Skiba et al. (2002) argue the case for no behavioral differences between Black and White children in referrals, but the most extreme reason for being given a referral ("threat") that distinguished between Black and White student referrals was significantly higher for Black students. Other study, e.g., Rocque and Paternoster (2011), collapses across severity in misconduct because of the infrequent nature of the more severe events, making it difficult to draw clear conclusions. In contrast, Lewis, Butler, Bonner, and Joubert (2010) found that Black boys compared to White boys were about twice as likely to engage in objective, severe behavior such as fighting with another student or making threats against another student. Regardless, the point for this paper is that experimental social psychology studies on this topic do not incorporate these distributional patterns into their designs or conclusions.

References

ACT. (2017). *The Condition of College & Career Readiness* 2017 (Tech. Rep.).

Akinola, M. N. (2009). *Deadly decisions: An examination of racial bias in the decision to shoot under threat*. Dissertation, Harvard University.

Andreoni, J., Kuhn, M. A., List, J. A., Samek, A., Sokal, K., & Sprenger, C. (2019). Toward an understanding of the development of time preferences: Evidence from field experiments. *Journal of Public Economics*, *177*, 104039.

Arcuri, L., Castelli, L., Galdi, S., Zogmaister, C., & Amadori, A. (2008). Predicting the vote: Implicit attitudes as predictors of the future behavior of decided and undecided voters. *Political Psychology*, *29*, 369–387.

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.

Banks, R. R., Eberhardt, J. L., & Ross, L. (2006). Discrimination and implicit bias in a racially unequal society. *California Law Review*, *94*, 1169–1190.

Barden, J., Maddux, W. W., Petty, R. E., & Brewer, M. B. (2004). Contextual moderation of racial bias: The impact of social roles on controlled and automatically activated attitudes. *Journal of Personality and Social Psychology*, *87*, 5–22.

Bargh, J. A. (1989). Conditional automaticity: Varieties of automatic influence in social perception and cognition. In J. S. Uleman & J. A. Bargh (Eds.), *Unintended thought* (pp. 3–51). Guilford.

Bargh, J. A. (1999). The cognitive monster: The case against the controllability of automatic stereotype effects. In S. Chaiken & Y. Trope (Eds.), *Dual process theories in social psychology* (pp. 361–382). Guilford.

Barnes, J., Jorgensen, C., Beaver, K. M., Boutwell, B. B., & Wright, J. P. (2015). Arrest prevalence in a national sample of adults: The role of sex and race/ethnicity. *American Journal of Criminal Justice*, *40*, 457–465.

Beaver, K. M., DeLisi, M., Wright, J. P., Boutwell, B. B., Barnes, J. C., & Vaughn, M. G. (2013). No evidence of racial discrimination in criminal justice processing: Results from the national longitudinal study of adolescent health. *Personality and Individual Differences*, *55*, 29–34.

Benbow, C. P., & Stanley, J. C. (1980). Sex differences in mathematical ability: Fact or artifact? *Science (New York, N.Y.)*, *210*, 1262–1264.

Benbow, C. P., Lubinski, D., Shea, D. L., & Eftekhari-Sanjani, H. (2000). Sex differences in mathematical reasoning ability at age 13: Their status 20 years later. *Psychological Science*, *11*, 474–480.

Bertrand, M., & Mullainathan, S. (2004). Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *American Economic Review*, *94*, 991–1013.

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

Blanton, H., & Jaccard, J. (2008). Unconscious racism: A concept in pursuit of a measure. *Annual Review of Sociology*, *34*, 277–297.

Blanton, H., Jaccard, J., Christie, C., & Gonzales, P. M. (2007). Plausible assumptions, questionable assumptions and post hoc rationalizations: Will the real IAT, please stand up? *Journal of Experimental Social Psychology*, *43*, 399–409.

Blanton, H., Jaccard, J., Gonzales, P. M., & Christie, C. (2006). Decoding the implicit association test: Implications for criterion prediction. *Journal of Experimental Social Psychology*, *42*, 192–212.

Blanton, H., Jaccard, J., Klick, J., Mellers, B., Mitchell, G., & Tetlock, P. E. (2009). Strong claims and weak evidence: Reassessing the predictive validity of the IAT. *Journal of Applied Psychology*, *94*, 567–582.

Blanton, H., Jaccard, J., Strauts, E., Mitchell, G., & Tetlock, P. E. (2015). Toward a meaningful metric of implicit prejudice. *Journal of Applied Psychology*, *100*, 1468–1481.

https://doi.org/10.1017/S0140525X21000017 Published online by Cambridge University Press

Byrnes, J. P., Miller, D. C., & Schafer, W. D. (1999). Gender differences in risk taking: A meta-analysis. *Psychological Bulletin*, *125*, 367–383.

Carey, B., & Goode, E. (2016). Police try to lower bias, but under pressure, it isn't so easy. *New York Times*. Retrieved from http://www.nytimes.com/2016/07/12/science/bias-reduction-programs.html.

Ceci, S. J., & Williams, W. M. (2010). Sex differences in math-intensive fields. *Current Directions in Psychological Science*, *19*, 275–279.

Ceci, S. J., & Williams, W. M. (2011). Understanding current causes of women's underrepresentation in science. *Proceedings of the National Academy of Sciences*, *108*, 3157–3162.

Ceci, S. J., Williams, W. M., & Barnett, S. M. (2009). Women's underrepresentation in science: Sociocultural and biological considerations. *Psychological Bulletin*, *135*, 218–261.

Cesario, J. (2021). On selective emphasis, broad agreement, and future directions: Reply to Ross, Winterhalder, & McElreath. Retrieved from psyarxiv.com/2p5eg.

Cesario, J., & Carrillo, A. (in press). Racial bias in police officer deadly force decisions: What has social cognition learned? In D. E. Carlston, K. Johnson, & K. Hugenberg (Eds.), *The Oxford handbook of social cognition* (2nd ed.). Oxford University Press.

Cesario, J., Johnson, D. J., & Terrill, W. (2019). Is there evidence of racial disparity in police use of deadly force? Analyses of officer-involved fatal shootings in 2015–2016. *Social Psychological and Personality Science*, *10*, 586–595.

Cheng, S. D. (2020). *Careers Versus Children: How Childcare Affects the Academic Tenure-Track Gender Gap*. Retrieved from https://scholar.harvard.edu/files/sdcheng/files/sdcheng_kids_jmpv7.pdf.

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.

Correll, J., Hudson, S. M., Guillermo, S., & Ma, D. S. (2014). The police officer's dilemma: A decade of research on racial bias in the decision to shoot. *Social and Personality Psychology Compass*, *8*, 201–213.

Correll, J., Park, B., Judd, C. M., & Wittenbrink, B. (2002). The police officer's dilemma: Using ethnicity to disambiguate potentially threatening individuals. *Journal of Personality and Social Psychology*, *83*, 1314–1329.

Correll, J., Park, B., Judd, C. M., Wittenbrink, B., Sadler, M. S., & Keesee, T. (2007). Across the thin blue line: Police officers and racial bias in the decision to shoot. *Journal of Personality and Social Psychology*, *92*, 1006–1023.

Correll, J., Wittenbrink, B., Park, B., Judd, C. M., & Goyle, A. (2011). Dangerous enough: Moderating racial bias with contextual threat cues. *Journal of Experimental Social Psychology*, *47*, 184–189.

Cortés, P., & Pan, J. (2020). Children and the remaining gender gaps in the labor market. *NBER Working Paper* 27980.

Cox, W. T. L., & Devine, P. G. (2016). Experimental research on shooter bias: Ready (or relevant) for application in the courtroom? *Journal of Applied Research in Memory and Cognition*, *5*, 236–238.

Cox, W. T. L., Devine, P. G., Plant, E. A., & Schwartz, L. L. (2014). Toward a comprehensive understanding of officers' shooting decisions: No simple answers to this complex problem. *Basic and Applied Social Psychology*, *36*, 356–364.

Cvencek, D., Greenwald, A. G., & Meltzoff, A. N. (2011a). Measuring implicit attitudes of 4-year-olds: The preschool implicit association test. *Journal of Experimental Child Psychology*, *109*, 187–200.

Cvencek, D., Meltzoff, A. N., & Greenwald, A. G. (2011b). Math–gender stereotypes in elementary school children. *Child Development*, *82*, 766–779.

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

Dasgupta, N., & Stout, J. G. (2012). Contemporary discrimination in the lab and field: Benefits and obstacles of full-cycle social psychology. *Journal of Social Issues*, *68*, 399–412.

Davis, E., Whyde, A., & Langton, L. (2018). Contacts between police and the public, 2015. *Bureau of Justice Statistics. U.S. Department of Justice*.

DeNavas-Walt, C., Proctor, B., & Smith, J. (2013). Income, poverty, and health insurance coverage in the United States: 2012. *U.S. Department of Commerce*.

Devine, P. G. (1989). Stereotypes and prejudice: Their automatic and controlled components. *Journal of Personality and Social Psychology*, *56*, 5–18.

Dhami, M. K., Hertwig, R., & Hoffrage, U. (2004). The role of representative design in an ecological approach to cognition. *Psychological Bulletin*, *130*, 959–988.

Diekman, A. B., Steinberg, M., Brown, E. R., Belanger, A. L., & Clark, E. K. (2017). A goal congruity model of role entry, engagement, and exit: Understanding communal goal processes in STEM gender gaps. *Personality and Social Psychology Review*, *21*, 142–175.

Dovidio, J. F. (2001). On the nature of contemporary prejudice: The third wave. *Journal of Social Issues*, *57*, 829–849.

Dovidio, J. F., & Gaertner, S. L. (2000). Aversive racism and selection decisions: 1989 and 1999. *Psychological Science*, *11*, 315–319.

Dovidio, J. F., Penner, L. A., Albrecht, T. L., Norton, W. E., Gaertner, S. L., & Shelton, J. N. (2008). Disparities and distrust: The implications of psychological processes for understanding racial disparities in health and health care. *Social Science & Medicine*, *67*, 478–486.

Dreifus, C. (2015). Perceptions of race at a glance. *New York Times*. Retrieved from http://www.nytimes.com/2015/01/06/science/a-macarthur-grant-winner-tries-to-unearth-biases-to-aid-criminal-justice.html.

Duncan, B. L. (1976). Differential social perception and attribution of intergroup violence: Testing the lower limits of stereotyping of blacks. *Journal of Personality and Social Psychology*, *34*, 590–598.

Eagly, A. H., & Johannesen-Schmidt, M. C. (2001). The leadership styles of women and men. *Journal of Social Issues*, *57*, 781–797.

Eagly, A. H., & Johnson, B. T. (1990). Gender and leadership style: A meta-analysis. *Psychological Bulletin*, *108*, 233–256.

Eagly, A. H., Nater, C., Miller, D. I., Kaufmann, M., & Sczesny, S. (2020). Gender stereotypes have changed: A cross-temporal meta-analysis of U.S. public opinion polls from 1946 to 2018. *American Psychologist*, *75*, 301–315.

Eagly, A. H., Wood, W., & Diekman, A. H. (2000). Social role theory of sex differences and similarities: A current appraisal. In T. Eckes & H. M. Trautner (Eds.), *The developmental social psychology of gender* (pp. 123–174). Erlbaum.

Fazio, R. H. (1990). Multiple processes by which attitudes guide behavior: The MODE model as an integrative framework. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 23, pp. 75–109). Elsevier.

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.

Felson, R. B. (1991). Blame analysis: Accounting for the behavior of protected groups. *The American Sociologist*, *22*, 5–23.

Fiedler, K., Messner, C., & Bluemke, M. (2006). Unresolved problems with the "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.

Fiske, S. T., & Neuberg, S. L. (1990). A continuum of impression formation, from category-based to individuating processes: Influences of information and motivation on attention and interpretation. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 23, pp. 1–74). Elsevier.

Fryer Jr, R. G. (2016). *An empirical analysis of racial differences in police use of force* (Tech. Rep.). National Bureau of Economic Research.

Fryer Jr, R. G., & Levitt, S. D. (2010). An empirical analysis of the gender gap in mathematics. *American Economic Journal: Applied Economics*, *2*, 210–240.

Fyfe, J. J. (1980). Geographic correlates of police shooting: A microanalysis. *Journal of Research in Crime and Delinquency*, *17*, 101–103.

Gawronski, B. (2019). Six lessons for a cogent science of implicit bias and its criticism. *Perspectives on Psychological Science*, *14*, 574–595.

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). The Guilford Press.

Geller, W. A., & Karales, K. J. (1981). Shootings of and by Chicago police: Uncommon crises – part I: Shootings by Chicago police. *Journal of Criminal Law & Criminology*, *72*, 1813–1866.

Gigerenzer, G., Hoffrage, U., & Kleinbölting, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review*, *98*, 506–528.

Gottfredson, L. S. (1997). Mainstream science on intelligence: An editorial with 52 signatories, history, and bibliography. *Intelligence*, *24*, 13–23.

Gottfredson, L. S. (1998). The general intelligence factor. *Scientific American Presents*, *9*, 24–29.

Gottfredson, L. S. (2004). Intelligence: Is it the epidemiologists' elusive "fundamental cause" of social class inequalities in health? *Journal of Personality and Social Psychology*, *86*, 174–199.

Gray, N. S., Brown, A. S., MacCulloch, M. J., Smith, J., & Snowden, R. J. (2005). An implicit test of the associations between children and sex in pedophiles. *Journal of Abnormal Psychology*, *114*, 304–308.

Green, A. R., Carney, D. R., Pallin, D. J., Ngo, L. H., Raymond, K. L., Iezzoni, L. I., & Banaji, M. R. (2007). Implicit bias among physicians and its prediction of thrombolysis decisions for Black and White patients. *Journal of General Internal Medicine*, *22*, 1231–1238.

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

Greenwald, A. G., & Krieger, L. H. (2006). Implicit bias: Scientific foundations. *California Law Review*, *94*, 945–967.

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

Haidt, J. (2011). The bright future of post-partisan social psychology. Talk given at the annual meeting of the Society for Personality and Social Psychology, San Antonio, TX. Retrieved from http://people.virginia.edu/~jdh6n/postpartisan.html.

https://doi.org/10.1017/S0140525X21000017 Published online by Cambridge University Press

Hakim, C. (2006). Women, careers, and work-life preferences. *British Journal of Guidance & Counselling*, *34*, 279–294.

Halpern, D. F., Benbow, C. P., Geary, D. C., Gur, R. C., Hyde, J. S., & Gernsbacher, M. A. (2007). The science of sex differences in science and mathematics. *Psychological Science in the Public Interest*, *8*, 1–51.

Heckman, J. J. (1998). Detecting discrimination. *Journal of Economic Perspectives*, *12*, 101–116.

Heriot, G. L., & Somin, A. (2018). The department of education's Obama-era initiative on racial disparities in school discipline: Wrong for students and teachers, wrong on the law. *Texas Review of Law and Politics*, *Forthcoming; San Diego Legal Studies Paper No.* 18-321.

Higgins, E. T. (1996). Knowledge activation: Accessibility, applicability, and salience. In E. T. Higgins & A. W. Kruglanski (Eds.), *Social psychology: Handbook of basic principles* (pp. 133–168). Guilford Press.

Higgins, E. T., & Bargh, J. A. (1987). Social cognition and social perception. *Annual Review of Psychology*, *38*, 369–425.

Hogarth, R. M. (1981). Beyond discrete biases: Functional and dysfunctional aspects of judgmental heuristics. *Psychological Bulletin*, *90*, 197–217.

Hsia, J. (1988). *Asian Americans in higher education and at work*. Lawrence Erlbaum Associates, Inc.

Hsin, A., & Xie, Y. (2014). Explaining Asian Americans' academic advantage over whites. *Proceedings of the National Academy of Sciences*, *111*, 8416–8421.

Inbar, Y., & Lammers, J. (2012). Political diversity in social and personality psychology. *Perspectives on Psychological Science*, *7*, 496–403.

Inn, A., Wheeler, A. C., & Sparling, C. L. (1977). The effects of suspect race and situation hazard on police officer shooting behavior. *Journal of Applied Social Psychology*, *7*, 27–37.

James, L., James, S. M., & Vila, B. J. (2016). The reverse racism effect: Are cops more hesitant to shoot black than white suspects? *Criminology & Public Policy*, *15*, 457–479.

James, L., Klinger, D., & Vila, B. (2014). Racial and ethnic bias in decisions to shoot seen through a stronger lens: Experimental results from high-fidelity laboratory simulations. *Journal of Experimental Criminology*, *10*, 323–340.

James, L., Vila, B., & Daratha, K. (2013). Results from experimental trials testing participant responses to White, Hispanic and Black suspects in high-fidelity deadly force judgment and decision-making simulations. *Journal of Experimental Criminology*, *9*, 189–212.

Jarvis, S. N., & Okonofua, J. A. (2020). School deferred: When bias affects school leaders. *Social Psychological and Personality Science*, *11*, 492–498.

Johnson, D. J., Cesario, J., & Pleskac, T. J. (2018). How prior information and police experience impact decisions to shoot. *Journal of Personality and Social Psychology*, *115*, 601–623.

Johnson, D. J., Hopwood, C. J., Cesario, J., & Pleskac, T. J. (2017). Advancing research on cognitive processes in social and personality psychology: A diffusion model primer. *Social Psychological and Personality Science*, *8*, 413–423.

Jost, J. T. (2019). The IAT is dead, long live the IAT: Context-sensitive measures of implicit attitudes are indispensable to social and political psychology. *Current Directions in Psychological Science*, *28*, 10–19.

Jost, J. T., Rudman, L. A., Blair, I. V., Carney, D. R., Dasgupta, N., Glaser, J., & Hardin, C. D. (2009). The existence of implicit bias is beyond reasonable doubt: A refutation of ideological and methodological objections and executive summary of ten studies that no manager should ignore. *Research in organizational behavior*, *29*, 39–69.

Jussim, L. (2012a). Liberal privilege in academic psychology and the social sciences: Commentary on Inbar & Lammers (2012). *Perspectives on Psychological Science*, *7*, 504–507.

Jussim, L. (2012b). *Social perception and social reality: Why accuracy dominates bias and self-fulfilling prophecy*. Oxford University Press.

Jussim, L., Cain, T. R., Crawford, J. T., Harber, K., & Cohen, F. (2009). The unbearable accuracy of stereotypes. In T. D. Nelson (Ed.), *Handbook of prejudice, stereotyping, and discrimination* (pp. 199–227). Psychology Press.

Jussim, L., Crawford, J., Anglin, S., Chambers, J., Stevens, S., & Cohen, F. (2015a). Stereotype accuracy: One of the largest and most replicable effects in all of social psychology. In T. D. Nelson (Ed.), *Handbook of prejudice, stereotyping, and discrimination* (pp. 31–63). Psychology Press.

Jussim, L., Crawford, J. T., Anglin, S. M., & Stevens, S. T. (2015b). Ideological bias in social psychological research. In J. P. Forgas, K. Fiedler, & W. D. Crano (Eds.), *Social psychology and politics* (pp. 107–126). Psychology Press.

Jussim, L., Crawford, J. T., & Rubinstein, R. S. (2015c). Stereotype (in)accuracy in perceptions of groups and individuals. *Current Directions in Psychological Science*, *24*, 490–497.

Kang, J., & Banaji, M. R. (2006). Fair measures: Behavioral realist revision of affirmative action. *California Law Review*, *94*, 1063–1018.

Kiefer, A. K., & Sekaquaptewa, D. (2007). Implicit stereotypes, gender identification, and math-related outcomes: A prospective study of female college students. *Psychological Science*, *18*, 13–18.

Klein, G. A. (1998). *Sources of power: How people make decisions*. MIT Press.

Kleven, H., Landais, C., Posch, J., Steinhauer, A., & Zweimüller, J. (2020). Do Family Policies Reduce Gender Inequality? Evidence from 60 Years of Policy Experimentation. *NBER Working Paper* (w28082).

Kleven, H., Landais, C., & Søgaard, J. E. (2019). Children and gender inequality: Evidence from Denmark. *American Economic Journal: Applied Economics*, *11*, 181–209.

Klinger, D., Rosenfeld, R., Isom, D., & Deckard, M. (2016). Race, crime, and the micro-ecology of deadly force. *Criminology & Public Policy*, *15*, 193–222.

Koch, A. J., D'Mello, S. D., & Sackett, P. R. (2015). A meta-analysis of gender stereotypes and bias in experimental simulations of employment decision making. *Journal of Applied Psychology*, *100*, 128–161.

Koenig, A. M., & Eagly, A. H. (2014). Evidence for the social role theory of stereotype content: Observations of groups' roles shape stereotypes. *Journal of Personality and Social Psychology*, *107*, 371–392.

Kristof, N. (2014). Is everyone a little bit racist? *New York Times*. Retrieved from http://www.nytimes.com/2014/08/28/opinion/nicholas-kristof-is-everyone-a-little-bit-racist.html.

Krueger, J., & Rothbart, M. (1988). Use of categorical and individuating information in making inferences about personality. *Journal of Personality and Social Psychology*, *55*, 187–195.

Kunda, Z., & Spencer, S. J. (2003). When do stereotypes come to mind and when do they color judgment? A goal-based theoretical framework for stereotype activation and application. *Psychological Bulletin*, *129*, 522–544.

Kunda, Z., & Thagard, P. (1996). Forming impressions from stereotypes, traits, and behaviors: A parallel-constraint-satisfaction theory. *Psychological Review*, *103*, 284–308.

Lee, K., & Ashton, M. C. (2020). Sex differences in HEXACO personality characteristics across countries and ethnicities. *Journal of Personality*, *88*, 1075–1090.

Levine, J. M., Resnick, L. B., & Higgins, E. T. (1993). Social foundations of cognition. *Annual Review of Psychology*, *44*, 585–612.

Lewis, C. W., Butler, B. R., Bonner III, F. A., & Joubert, M. (2010). African American male discipline patterns and school district responses resulting impact on academic achievement: Implications for urban educators and policy makers. *Journal of African American Males in Education*, *1*, 7–25.

Lhamon, C., & Samuels, J. (2014). Dear colleague letter on the nondiscriminatory administration of school discipline. *Washington, DC: U.S. Department of Education, Office of Civil Rights & U.S. Department of Justice, Civil Rights Division*.

Lick, D. J., Alter, A. L., & Freeman, J. B. (2018). Superior pattern detectors efficiently learn, activate, apply, and update social stereotypes. *Journal of Experimental Psychology: General*, *147*, 209–227.

Lippa, R. (1998). Gender-related individual differences and the structure of vocational interests: The importance of the people–things dimension. *Journal of Personality and Social Psychology*, *74*, 996–1009.

Locksley, A., Borgida, E., Brekke, N., & Hepburn, C. (1980). Sex stereotypes and social judgment. *Journal of Personality and Social Psychology*, *39*, 821–831.

Logan, G. D. (2018). Automatic control: How experts act without thinking. *Psychological Review*, *125*, 453–485.

Lopez, G. (2017). Police shootings and brutality in the US: 9 things you should know. *Vox*. Retrieved from https://www.vox.com/cards/police-brutality-shootings-us/us-police-racism.

Loughlin, J. K., & Flora, K. (2017). *Shots fired: The misunderstandings, misconceptions, and myths about police shootings*. Simon and Schuster.

Lu, J. G., Nisbett, R. E., & Morris, M. W. (2020). Why East Asians but not South Asians are underrepresented in leadership positions in the United States. *Proceedings of the National Academy of Sciences*, *117*, 4590–4600.

Lubinski, D., & Benbow, C. P. (1992). Gender differences in abilities and preferences among the gifted: Implications for the math-science pipeline. *Current Directions in Psychological Science*, *1*, 61–66.

Lynn, R. (2004). The intelligence of American Jews. *Personality and Individual Differences*, *36*, 201–206.

Lynn, R., & Irwing, P. (2004). Sex differences on the progressive matrices: A meta-analysis. *Intelligence*, *32*, 481–498.

Ma, D. S., & Correll, J. (2011). Target prototypicality moderates racial bias in the decision to shoot. *Journal of Experimental Social Psychology*, *47*, 391–396.

Ma, D., Graves, S., & Alvarado, J. (2019). A spatial analysis of officer-involved shootings in Los Angeles. *Yearbook of the Association of Pacific Coast Geographers*, *81*, 158–181.

Madon, S., Jussim, L., Guyll, M., Nofziger, H., Salib, E., Willard, J., & Scherr, K. C. (2018). The accumulation of stereotype-based self-fulfilling prophecies. *Journal of Personality and Social Psychology*, *115*(5), 825–844.

Martell, R. F., Lane, D. M., & Emrich, C. (1996). Male–female differences: A computer simulation. *American Psychologist*, *51*, 157–158.

McCauley, C., Stitt, C. L., & Segal, M. (1980). Stereotyping: From prejudice to prediction. *Psychological Bulletin*, *87*, 195–208.

McLanahan, S., & Percheski, C. (2008). Family structure and the reproduction of inequalities. *Annual Review Sociology*, *34*, 257–276.

Mekawi, Y., & Bresin, K. (2015). Is the evidence from racial bias shooting task studies a smoking gun? Results from a meta-analysis. *Journal of Experimental Social Psychology*, *61*, 120–130.

Mentch, L. (2020). On racial disparities in recent fatal police shootings. *Statistics and Public Policy*, *7*, 9–18.

https://doi.org/10.1017/S0140525X21000017 Published online by Cambridge University Press

Miller, A. L. (2019). Expertise fails to attenuate gendered biases in judicial decision-making. *Social Psychological and Personality Science*, *10*, 227–234.
Miller, T. R., Lawrence, B. A., Carlson, N. N., Hendrie, D., Randall, S., Rockett, I. R., & Spicer, R. S. (2017). Perils of police action: A cautionary tale from US data sets. *Injury Prevention*, *23*, 27–32.
Mitchell, G. (2012). Revisiting truth or triviality: The external validity of research in the psychological laboratory. *Perspectives on Psychological Science*, *7*, 109–117.
Mitchell, G. (2018). Jumping to conclusions: Advocacy and application of psychological research. In J. Crawford & L. Jussim (Eds.), *The politics of social psychology* (pp. 139–155). Routledge.
Mook, D. G. (1983). In defense of external invalidity. *American Psychologist*, *38*, 379–387.
Morrison, G. (2006). Police department and instructor perspectives on pre-service firearm and deadly force training. *Policing: An International Journal of Police Strategies & Management*, *29*, 226–245.
Mortensen, C. R., & Cialdini, R. B. (2010). Full-cycle social psychology for theory and application. *Social and Personality Psychology Compass*, *4*, 53–63.
Moss-Racusin, C. A., Dovidio, J. F., Brescoll, V. L., Graham, M. J., & Handelsman, J. (2012). Science faculty's subtle gender biases favor male students. *Proceedings of the National Academy of Sciences*, *109*, 16474–9.
Musu-Gillette, L., Zhang, A., Wang, K., Zhang, J., Kemp, J., Diliberti, M., & Oudekerk, B. A. (2018). *Indicators of school crime and safety: 2017*. National Children's Advocacy Center.
Neal, D. A., & Johnson, W. R. (1996). The role of premarket factors in Black–White wage differences. *Journal of Political Economy*, *104*, 869–895.
Nix, J., Campbell, B. A., Byers, E. H., & Alpert, G. P. (2017). A bird's eye view of civilians killed by police in 2015: Further evidence of implicit bias. *Criminology & Public Policy*, *16*, 309–340.
Nock, M. K., & Banaji, M. R. (2007). Prediction of suicide ideation and attempts among adolescents using a brief performance-based test. *Journal of Consulting and Clinical Psychology*, *75*, 707–715.
Nosek, B. A., Banaji, M. R., & Greenwald, A. G. (2002). Math = male, me = female, therefore math ≠ me. *Journal of Personality and Social Psychology*, *83*(1), 44–59.
Nosek, B. A., Greenwald, A. G., & Banaji, M. R. (2007). The implicit association test at age 7: A methodological and conceptual review. *Automatic Processes in Social Thinking and Behavior*, *4*, 265–292.
Nosek, B. A., & Smyth, F. L. (2011). Implicit social cognitions predict sex differences in math engagement and achievement. *American Educational Research Journal*, *48*, 1125–1156.
Nosek, B. A., Smyth, F. L., Sriram, N., Lindner, N. M., Devos, T., Ayala, A., … Greenwald, A. G. (2009). National differences in gender-science stereotypes predict national sex differences in science and math achievement. *Proceedings of the National Academy of Sciences*, *106*, 10593–7.
Okonofua, J. A., & Eberhardt, J. L. (2015). Two strikes: Race and the disciplining of young students. *Psychological Science*, *26*, 617–624.
Okonofua, J. A., Walton, G. M., & Eberhardt, J. L. (2016). A vicious cycle: A social–psychological account of extreme racial disparities in school discipline. *Perspectives on Psychological Science*, *11*, 381–398.
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.
Oswald, F. L., Mitchell, G., Blanton, H., Jaccard, J., & Tetlock, P. E. (2015). Using the IAT to predict ethnic and racial discrimination: Small effect sizes of unknown societal significance. *Journal of Personality and Social Psychology*, *108*, 562–571.
Palfai, T. P., & Ostafin, B. D. (2003). Alcohol-related motivational tendencies in hazardous drinkers: Assessing implicit response tendencies using the modified-IAT. *Behaviour Research and Therapy*, *41*, 1149–1162.
Petty, R. E., & Wegener, D. T. (1999). The elaboration likelihood model; current status and controversies. In S. Chaiken & Y. Trope (Eds.), *Dual process theories in social psychology* (pp. 41–72). Guilford.
Plant, E. A., Goplen, J., & Kunstman, J. W. (2011). Selective responses to threat: The roles of race and gender in decisions to shoot. *Personality and Social Psychology Bulletin*, *37*, 1274–1281.
Plant, E. A., & Peruche, B. M. (2005). The consequences of race for police officers' responses to criminal suspects. *Psychological Science*, *16*, 180–183. doi: 10.1111/j.0956-7976.2005.00800.x.
Platt, J. R. (1964). Strong inference. *Science (New York, N.Y.)*, *146*, 347–353.
Pleskac, T. J., Cesario, J., & Johnson, D. J. (2018). How race affects evidence accumulation during the decision to shoot. *Psychonomic Bulletin & Review*, *25*, 1301–1330. https://doi.org/10.3758/s13423-017-1369-6.
Pleskac, T. J., & Hertwig, R. (2014). Ecologically rational choice and the structure of the environment. *Journal of Experimental Psychology: General*, *143*, 2000–2019.
Pleskac, T. J., Johnson, D. J., Cesario, J., Terrill, W., & Gagnon, G. (under review). Modeling police officers' deadly force decisions in an immersive shooting simulator.
Price-Williams, D. R., & Ramirez, M. I. I. I. (1974). Ethnic differences in delay of gratification. *The Journal of Social Psychology*, *93*, 23–30.
Ring, K. (1967). Experimental social psychology: Some sober questions about some frivolous values. *Journal of Experimental Social Psychology*, *3*, 113–123.
Rocque, M., & Paternoster, R. (2011). Understanding the antecedents of the "school-to-jail" link: The relationship between race and school discipline. *The Journal of Criminal Law and Criminology*, *101*, 633–665.
Rooth, D. (2007). Implicit discrimination in hiring: Real world evidence (IZA Discussion Paper No. 2764). Bonn, Germany: Forschungsinstitut zur Zukunft der Arbeit (Institute for the Study of Labor).
Ross, C. T., Winterhalder, B., & McElreath, R. (2021). Racial disparities in police use of deadly force against unarmed individuals persist after appropriately benchmarking shooting data on violent crime rates. *Social Psychological and Personality Science*, *12*, 323–332.
Roth, P. L., Bevier, C. A., Bobko, P., Switzer III, F. S., & Tyler, P. (2001). Ethnic group differences in cognitive ability in employment and educational settings: A meta-analysis. *Personnel Psychology*, *54*, 297–330.
Rozin, P. (2009). What kind of empirical research should we publish, fund, and reward?: A different perspective. *Perspectives on Psychological Science*, *4*, 435–439.
Rudman, L. A., & Ashmore, R. D. (2007). Discrimination and the implicit association test. *Group Processes & Intergroup Relations*, *10*, 359–372.
Rudman, L. A., & Glick, P. (2001). Prescriptive gender stereotypes and backlash toward agentic women. *Journal of Social Issues*, *57*, 743–762.
Schimmack, U. (2020). The implicit association test: A method in search of a construct. *Perspectives on Psychological Science*, *16*, 396–414.
Scott, K., Ma, D. S., Sadler, M. S., & Correll, J. (2017). A social scientific approach toward understanding racial disparities in police shooting: Data from the department of justice (1980–2000). *Journal of Social Issues*, *73*, 701–722.
Shjarback, J. A., & Nix, J. (2020). Considering violence against police by citizen race/ethnicity to contextualize representation in officer-involved shootings. *Journal of Criminal Justice*, *66*, 101653-63.
Sim, J. J., Correll, J., & Sadler, M. S. (2013). Understanding police and expert performance: When training attenuates (vs. exacerbates) stereotypic bias in the decision to shoot. *Personality and Social Psychology Bulletin*, *39*, 291–304. doi: 10.1177/0146167212473157.
Skiba, R. J., Michael, R. S., Nardo, A. C., & Peterson, R. L. (2002). The color of discipline: Sources of racial and gender disproportionality in school punishment. *The Urban Review*, *34*, 317–342.
Smith, E. R., & DeCoster, J. (2000). Dual-process models in social and cognitive psychology: Conceptual integration and links to underlying memory systems. *Personality and Social Psychology Review*, *4*, 108–131.
Sowell, T. (2005). *Black rednecks and white liberals*. Encounter Books.
Sowell, T. (2008). *Discrimination and disparities*. Basic Books.
Srivastava, S. (2016). Everything is fucked: The syllabus [Blog post]. Retrieved from https://thehardestscience.com/2016/08/11/everything-is-fucked-the-syllabus/.
Stanley, D. A., Sokol-Hessner, P., Banaji, M. R., & Phelps, E. A. (2011). Implicit race attitudes predict trustworthiness judgments and economic trust decisions. *Proceedings of the National Academy of Sciences*, *108*, 7710–7715.
Stickle, B. (2016). A national examination of the effect of education, training and pre-employment screening on law enforcement use of force. *Justice Policy Journal*, *13*, 1–15.
Su, R., & Rounds, J. (2015). All STEM fields are not created equal: People and things interests explain gender disparities across STEM fields. *Frontiers in Psychology*, *6*, 1–20.
Su, R., Rounds, J., & Armstrong, P. I. (2009). Men and things, women and people: A meta-analysis of sex differences in interests. *Psychological Bulletin*, *135*, 859–884.
Taylor, A. (2011). The influence of target race on split-second shooting decisions in simulated scenarios :a Canadian perspective. (Doctor of Philosophy, Carleton University, Ottawa, Ontario). doi: 10.22215/etd/2011-09571.
Tetlock, P. E. (1985). Accountability: The neglected social context of judgment and choice. *Research in Organizational Behavior*, *7*, 297–332.
Tregle, B., Nix, J., & Alpert, G. P. (2019). Disparity does not mean bias: Making sense of observed racial disparities in fatal officer-involved shootings with multiple benchmarks. *Journal of Crime and Justice*, *42*, 18–31.
Uhlmann, E. L., Brescoll, V. L., & Paluck, E. L. (2006). Are members of low status groups perceived as bad, or badly off? Egalitarian negative associations and automatic prejudice. *Journal of Experimental Social Psychology*, *42*, 491–499.
Valla, J. M., Ceci, S. J. (2014). Breadth-based models of women's underrepresentation in STEM fields: An integrative commentary on Schmidt (2011) and Nye et al. (2012). *Perspectives on Psychological Science*, *9*, 219–224.
von Hippel, W., Brener, L., & von Hippel, C. (2008). Implicit prejudice toward injecting drug users predicts intentions to change jobs among drug and alcohol nurses. *Psychological Science*, *19*, 7–11.
Wheeler, A. P., Phillips, S. W., Worrall, J. L., & Bishopp, S. A. (2017). What factors influence an officer's decision to shoot? The promise and limitations of using public data. *Justice Research and Policy*, *18*, 48–76. doi: 10.1177/1525107118759900.

https://doi.org/10.1017/S0140525X21000017 Published online by Cambridge University Press

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.

Worrall, J. L., Bishopp, S. A., Zinser, S. C., Wheeler, A. P., & Phillips, S. W. (2018). Exploring bias in police shooting decisions with real shoot/don't shoot cases. *Crime Delinquency*, *64*, 1171–1192. doi: 10.1177/0011128718756038.

Wright, J. P., Morgan, M. A., Coyne, M. A., Beaver, K. M., & Barnes, J. (2014). Prior problem behavior accounts for the racial gap in school suspensions. *Journal of Criminal Justice*, *42*, 257–266.

Zytkoskee, A., Strickland, B. R., & Watson, J. (1971). Delay of gratification and internal versus external control among adolescents of low socioeconomic status. *Developmental Psychology*, *4*, 93–98.

Open Peer Commentary

Practical consequences of flawed social psychological research on bias

Hal R. Arkes

Department of Psychology, Ohio State University, Columbus, OH 43210, USA.
arkes.1@osu.edu
https://psychology.osu.edu/people/arkes.1

doi:10.1017/S0140525X21000649, e67

Abstract

The flaws in social psychological research pointed out by Cesario have societal costs. These include ignoring crucial base rates thereby degrading the effectiveness of policy decisions, generalizing the conclusions derived from experiments on non-professionals thereby distorting the public's view of professional law enforcement personnel, questionable accusations of racism, and mis-attributions of the causes of racial differences in behavior.

Cesario points out that the conditions in social psychological experiments that foster the manifestation of bias are largely absent in "real-world" domains. Nevertheless, the bias detected in such flawed experiments has been deemed "widespread" and "pervasive" in "real-world" domains. The serious flaws Cesario has identified have not impeded the overselling of this "bias." In fact, they have facilitated this overselling, some of whose costs I will now enumerate.

First, as Cesario points out, base rates of criminality, school rule violations, and other important data are not included in the stimulus materials used in social psychology experiments. People, therefore, cannot use these data, and resulting judgments necessarily don't correspond to the judgments made in the "real world" where such information is often available. When people are given the opportunity to use base rates in social psychological experiments on race, they are deemed to be "Bayesian bigots" (Banaji, 2003) and are perceived by observers as "unintelligent" reasoners (Cao, Kleiman-Weiner, & Banaji, 2019). Should people ignore base rates in their inter-racial social judgments? The underutilization of base rates in criminal investigations has been shown to cost approximately 1,900 lives of minority members per year (Farmer & Terrell, 2001). So which is best: (a) ignoring base rates, policing every racial group precisely equally, and annually losing minority lives to homicide, or (b) honoring base rates with more police in minority as opposed to majority neighborhoods and saving lives? Deeming Bayesian reasoners to be bigots has a serious societal cost. In the judgment and decision-making literature not using relevant base rates is considered to be an error in reasoning (Kahneman & Tversky, 1973). In the social psychology literature, not providing relevant base rates for people to use in their judgments is considered to be appropriate.

Second, as Cesario points out, the shooter bias studies typically provide neither history of nor prior interaction with the suspect, and such studies often use laypersons rather than trained professionals. Manifestations of bias in such unrealistic situations help to fuel calls to defund the police who are accused of harboring implicit bias in realistic situations. Recent analyses reveal neither racial differences in the use of extreme force on the part of police (Fryer, 2019) nor disproportionate arrests of Blacks (Beck, 2021). These results contrast sharply with the results of shooter bias studies using non-professionals in unrealistic situations. A recent poll in minority neighborhoods found that over 80% of residents in those neighborhoods want either an increase in police presence or no decrease in police presence (Grzeszczak, 2020). Police reticence, resignations, and retirements have resulted in increased crime in many of these neighborhoods during the last few years (e.g., Lauritien, 2021). Debilitating the police because of their supposed "widespread" and "pervasive" bias has a serious cost.

Third, the conclusions drawn from social psychological research have prompted charges of racism even when the empirical evidence provides no support for such accusations. To cite one example, Sowell (2019, p. 89) points out that in 2000 the U.S. Civil Rights Commission, on which I was a state advisory committee member, reported that 44.6% of Black applicants were denied a mortgage, but only 22.3% of Whites were denied. This led to a chorus of demands that the government should crack down on this level of abject discrimination. Cesario discussed several laboratory studies in the business domain in which discrimination was the purported motivation for such racial differences. However, further inspection of the mortgage data showed that the credit scores for the Black applicants were lower than the credit scores for Whites. Because the lending sources had "skin in the game" – the money that they would loan to the mortgage applicants – they were prudent to base their mortgage decisions on the financial qualifications of the applicant. This example illustrates the role of incentives. In social psychology laboratory studies, there is no incentive for either discriminatory or non-discriminatory behavior. In contrast, a bigoted mortgage lender who refuses to lend money to qualified Black applicants will suffer negative financial consequences. In the "real world" prejudice can have a cost. Thus, it will be less likely to be manifested than in a social psychology experiment in which bigotry goes unpunished.

Cesario points out that accusations of prejudice may occur in such instances in which a teacher might interpret a child's behavior differently depending on the race of the child. Unfortunately, there is no perfectly objective way of categorizing the child's behavior. The police face this ambiguous situation frequently in which mis-categorizing a behavior can have disastrous consequences. Fortunately, there are some instances in which the categorization of the behavior is certain and thus where prejudice can be accurately discerned. An example occurred in New Jersey when the police were criticized for giving more speeding tickets to African American drivers than White drivers. A research project was initiated in which a camera took pictures of drivers on the New Jersey turnpike while a radar gun measured each driver's speed. A trio of three persons looked at a still photograph of

each car's driver, and two of the three evaluators had to agree on what the race of the driver was. Because the trio was looking at a still photograph, they could not determine if this car was exceeding the speed limit. "Speeding" was defined as traveling at least 15 miles per hour over the speed limit. The data showed that the drivers identified as African American were nearly twice as likely to be speeding as the drivers identified as White. In other words, the differential number of speeding tickets was not because of "implicit prejudice" (Hinnant, 2002). Of course, laboratory experiments are unlikely to contain spontaneous levels of natural behavior which can be categorized with certainty. Thus, it is risky to generalize racial differences detected in laboratory experiments to societal-level racial differences.

Social psychology experiments can be valuable and informative. However, their generalization to societal-level issues must be done with humility, better design, and much more attention to external validity.

Financial support. This research received no specific grant from any funding agency, commercial, or not-for-profit sectors.

Conflict of interest. None.

References

Banaji, M. R. (2003). *Mind bugs: The psychology of ordinary prejudice*. Colloquium presentation at The Ohio State University, Columbus.

Beck, A. J. (2021). *Race and ethnicity of violent crime offenders and arrestees, 2018*. United States Department of Justice, Office of Justice Programs, Bureau of Justice Statistics, NCJ #25569.

Cao, J., Kleiman-Weiner, M., & Banaji, M. R. (2019). People make the same Bayesian judgment they criticize in others. *Psychological Science*, *30*(1), 20–31. doi:10.1177/0956797618805750

Farmer, A., & Terrell, D. (2001). Crime versus justice: Is there a trade-off? *Journal of Law and Economics*, *44*, 345–366.

Fryer, Jr., R. G. (2019) An empirical analysis of racial differences in police use of force. *Journal of Political Economy*, *127* 1210–1261. https://doi.org/10.1086/701423.

Grzeszczak, J. (2020). 81% of Black Americans Don't Want Less Police Presence Despite Protests – Some Want More Cops: Poll. *Newsweek.com*. https://advance-lexis-com.proxy.lib.ohio-state.edu/api/document?collection=news&id=urn:contentItem:60HK-Y0V1-JBR6-928Y-00000-00&context=1516831.

Hinnant, L. (2002). New Jersey state officials release results of controversial racial speeding study. https://advance-lexis-com.proxy.lib.ohio-state.edu/api/document?collection=news&id=urn:contentItem:45FG-FNT0-009F-R0KT-00000-00&context=1516831.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, *80*, 237–251.

Lauritien, J. (2021). Early Minneapolis crime statistics show 250% increase in gunshot victims. https://minnesota.cbslocal.com/2021/01/22/ early-minneapolis-crime-statistics-show-250%- increase-in gunshot-victims/.

Sowell, T. (2019). *Discrimination and disparities*. Hachette Book Group, Inc.

Social bias insights concern judgments rather than real-world decisions

Michał Białek[a] and Igor Grossmann[b]

[a]Institute of Psychology, Faculty of Historical and Pedagogical Sciences, University of Wrocław, 50-529 Wrocław, Poland and [b]Department of Psychology, University of Waterloo, Waterloo, ON N2L 3G1, Canada.
michal.bialek3@uwr.edu.pl, igor.grossmann@uwaterloo.ca

doi:10.1017/S0140525X21000728, e68

Abstract

Judgments differ from decisions. Judgments are more abstract, decontextualized, and bear fewer consequences for the agent. In pursuit of experimental control, psychological experiments on bias create a simplified, bare-bone representation of social behavior. These experiments resemble conditions in which people judge others, but not how they make real-world decisions.

During a family dinner, a chat sways into politics. Most guests have an opinion about politicians and their conduct, labeling some politicians as irrational or immoral. How did the guests form these opinions? Unless the family belongs to political elites, these opinions are likely based on news reports or commentaries on social media. Such news posts are brief, decontextualized, and ill-defined, with uncertainty about politicians' interests and preferences – caricatured abstractions, often far off from reality. However, even if guests are aware of these limits to their knowledge, it hardly stops them from expressing rather strong opinions about politicians.

We use politicians, but many everyday judgments can be characterized this way – fast and based on limited information. Another feature of judging others is that such judgments – as long as they don't trigger a response – typically bear little consequence to the person expressing them.

In contrast, consider how guests at the same family dinner would decide on particular policies. Here, much more information is likely to be incorporated in their decision: Goals and preferences are concrete, and the decision is put in a rich context of their political and social background. Notably, the consequences of their decisions will have a big real-world impact via the outcomes of their choice. Here, the decision about policies appears qualitatively different from the judgment about politicians proposing these policies, even though both judgment and decision-making processes can inform each other (Ariely & Norton, 2008).

To further unpack this distinction between decision and judgment, consider a visit to a restaurant. You notice pulled pork and smoked ribs on the menu. Though both are your favorites, you like pulled pork a bit more, say 8 versus 9 points on a 10-point scale. When ranking the two meals, you will always put pulled pork before smoked ribs. But this does not mean you will always order pulled pork and will never order ribs. There is more for you to consider when deciding about a particular order than merely the judgment in a form of ranking your preferences. Judgments tell us about values, beliefs, and preferences, whereas decisions also tell us how judgments are (mis)applied into action. The strength of external factors in judgment and decision making is illustrated by the decoy effect – that is, a phenomenon whereby people who value option A over B may sometimes chose option B when presented with a third option C (Huber, Payne, & Puto, 1982). Note, we do not claim decision is just distorted judgment. We simply highlight that judgments and decisions differ in several ways.

These insights have direct implications to social bias research. In a typical study, participants are presented with experimental stimuli (words and pictures) and are asked to choose, often in a narrow time slot, between the two options. For example, a picture of a black or white man is presented with an item, and participants are asked to quickly choose to "shoot" or not to. The experimental set-up aims to create a simplified, bare-bone representation of behavior. For ethical reasons, other experimental paradigms may

not always be possible. In pursuit of experimental control, these experiments are stripped of much context information, and consequently diverge from real-world decisions. Cesario highlights these differences, raising a provocative question about the reduced utility of such experiments. When considering predictions of real-world behavior, we agree with this sentiment. At the same time, we note that such decontextualized studies may, in fact, resemble conditions in which judgments are made. As we outlined in examples above, judgments are often decontextualized, are made on the fly, and bear little consequences to the agent. Thus, social bias experiments may well mimics conditions in which people make judgments, especially judgments of the conduct of others, despite lack of predictive power for the real-world decisions.

If judgments and decisions are two different but interrelated phenomena, there are implications for how to interpret the results of social bias experiments. Cesario notes that having observed an implicit bias in an agent, one cannot infer that this agent will necessarily make biased real-life decisions. More likely, such person will make biased judgments.

How people make judgments is not only relevant to the study of social bias, but also for other areas of psychology, including the closely related research on moral psychology. Humans judge each other's morality at a much greater rate that they make morally charged decisions. Hence, learning about processes underlying moral judgments seems at least as important for moral psychology to answer than processes underlying moral decisions (Białek, Turpin, & Fugelsang, 2019). In a similar vein, people can judge others daily, but only sometimes actively discriminate against others. Although the latter may have more severe consequences, biased judgment can also be devastating. For example, biased judgment can support system justification (Jost, Banaji, & Nosek, 2004), and maintenance of discriminatory laws harming the group one is biased against. Bias can lead disadvantages groups to favor privileged outgroup (dos Santos & Pereira, 2021). Hence, discovering how biased judgments are formed allows us to understand psychological reasons for support of existing inequalities.

We believe that these criticized experiments on bias have an enormous value for social sciences: They inform us about core beliefs and preferences of particular social groups. Whether a person belonging to a given group will act on these beliefs is a distinct, and arguably more complex question. It may be unreasonable to expect researchers to comprehensively answer both in one project. After all, social scientists are not (well) trained in prediction modeling of social issues (Hofman, Sharma, & Watts, 2017; Yarkoni & Westfall, 2017), often ignore the broader cultural (Henrich, Heine, & Norenzayan, 2010) and cross-temporal factors (Grossmann & Varnum, 2015; Varnum & Grossmann, 2017), and consequently appear as inaccurate in their assessments of broader societal issues as an average person on the street (Hutcherson et al., 2021).

Instead, it may be prudent to establish robust scientific evidence one step at a time. For instance, researchers may start by focusing on the study of agent's preferences and biases, prior to scaling up a model to predict real-life decisions. The second step will require extending the experiments by considering contextual cues. The third step will benefit from greater integration of insights from computational social sciences and complex systems for predictive modeling of human behavior (e.g., Hofman et al., 2017; Yarkoni & Westfall, 2017). Only by integrating these steps together, social scientists can start translating insights from experiments about social judgment biases into the study of real-world behavior.

Financial support. The current project was financed by the resources of Polish National Science Centre (NCN) assigned by the decision no. 2017/26/D/HS6/01159 to MB. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Conflict of interest. None.

References

Ariely, D., & Norton, M. I. (2008). How actions create – Not just reveal – Preferences. *Trends in Cognitive Sciences*, *12*(1), 13–16. https://doi.org/10.1016/j.tics.2007.10.008.

Białek, M., Turpin, M. H., & Fugelsang, J. A. (2019). What is the right question for moral psychology to answer? Commentary on Bostyn, Sevenhant, and Roets (2018). *Psychological Science*, *30*(9), 1383–1385. https://doi.org/10.1177/0956797618815171.

dos Santos, M. F., & Pereira, C. R. (2021). The social psychology of a selective national inferiority complex: Reconciling positive distinctiveness and system justification. *Journal of Experimental Social Psychology*, *95*, 104118. https://doi.org/10.1016/j.jesp.2021.104118.

Grossmann, I., & Varnum, M. E. W. (2015). Social structure, infectious diseases, disasters, secularism, and cultural change in America. *Psychological Science*, *26*(3), 311–324. https://doi.org/10.1177/0956797614563765.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). Beyond WEIRD: Towards a broad-based behavioral science. *Behavioral and Brain Sciences*, *33*(2–3), 111–135. https://doi.org/10.1017/S0140525X10000725.

Hofman, J. M., Sharma, A., & Watts, D. J. (2017). Prediction and explanation in social systems. *Science (New York, N.Y.)*, *355*(6324), 486–488. https://doi.org/10.1126/science.aal3856.

Huber, J., Payne, J. W., & Puto, C. (1982). Adding asymmetrically dominated alternatives: Violations of regularity and the similarity hypothesis. *Journal of Consumer Research*, *9*(1), 90. https://doi.org/10.1086/208899.

Hutcherson, C., Sharpinskyi, K., Varnum, M. E. W., Rotella, A., Wormley, A., Tay, L., & Grossmann, I. (2021). *Behavioral scientists and laypeople misestimate societal effects of COVID-19*. https://doi.org/10.31234/osf.io/g8f9s.

Jost, J. T., Banaji, M. R., & Nosek, B. A. (2004). A decade of system justification theory: Accumulated evidence of conscious and unconscious bolstering of the status quo. *Political Psychology*, *25*(6), 881–919. https://doi.org/10.1111/j.1467-9221.2004.00402.x.

Varnum, M. E. W., & Grossmann, I. (2017). Cultural change: The how and the why. *Perspectives on Psychological Science*, *12*(6), 956–972. https://doi.org/10.1177/1745691617699971.

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, *12*(6), 1100–1122. https://doi.org/10.1177/1745691617693393.

The missing consequences: A fourth flaw of experiments

Adam Thomas Biggs

Naval Special Warfare Command, Coronado, CA 92155, USA.
adam.t.biggs.mil@socom.mil

doi:10.1017/S0140525X21000686, e69

Abstract

Decisions are affected by the potential consequences as much as any factor during the decision-making process. This prospective influence represents another flaw overlooked by most experiments that raises questions about the use of certain laboratory paradigms. Lethal force encounters are a prime example of this problem, where negative consequences of slow decisions and wrong decisions should be considered alongside behavior.

The article well describes three flaws in experimental designs typically used to examine biases. However, there is another fundamental flaw that has been overlooked: *consequences*. Poor

performance has implications in real-world tasks where both slow and inaccurate responses lead to disastrous outcomes. The three flaws address each problem as it exists in a laboratory setting, yet there is no discussion regarding how the consequences of a decision might influence behavior. Specifically, your behavior can change when you know the situation is real and not just a simulation.

Nowhere is this omission more salient than in the first-person shooter paradigm. The author discusses dispatch priming (Taylor, 2020), contextual information about the environment (Correll, Wittenbrink, Park, Judd, & Goyle, 2011), realism through enhanced simulator scenarios (James, Klinger, & Vila, 2014; James, Vila, & Daratha, 2013), and even a thorough discussion about training (Cesario & Carrillo, in press; Cox & Devine, 2016; Sim, Correll, & Sadler, 2013). Consequences remain conspicuously absent. For example, laboratory-based paradigms do not regularly impose consequences after making a poor decision. Shooting tasks rarely impose any penalty for firing upon an unarmed person, and when they do, the consequence is more likely to be a point-based deduction (Biggs, Cain, & Mitroff, 2015). Experiments normally just proceed to the next trial, whereas these real-world errors are followed by detailed officer-involved shooting investigations and sometimes criminal punishment.

There are also no consequences to the shooter for moving too slowly. For all the deliberation about experimental paradigms, there is no discussion about a hazard present in many lethal force encounters – hostiles can shoot back. Realistic lethal force engagements carry life-or-death significance for the shooter too as moving too slowly could mean being shot by a hostile adversary. This threat imposes consequences for failing to act in addition to the consequences for making the wrong decision. There is an entire literature on this topic absent from the discussion that thoroughly addresses the stress and anxiety present in lethal force scenarios because of pressure and consequence (Nieuwenhuys & Oudejans, 2010, 2011; Oudejans, 2008; Patton & Gamble, 2016). Among the various influences that might alter performance in a shooting task, hostile action should be represented with the same prominence and concern as using a realistic weapon. Moreover, consequences do not readily fit into any of the missing categories, which is why they should be described as a fourth flaw.

Contrast this prospective influence due to a course of action with the stated three flaws. For the missing information flaw, the concern is creating unrealistic conditions in the laboratory for the sake of experimental control. This flaw emphasizes making artificial factors more authentic, albeit the unintended consequence might be embracing superficially related components as more genuine when they are not as interrelated. One such instance is how marksmanship and decision-making are more disconnected than they seem, making marksmanship an orthogonal factor to the lethal force decision-making process (Blacker, Pettijohn, Roush, & Biggs, 2021). The irony is that authentic grip and firearm functions suffice for realism without being concerned about what the bullet does when it hits the target.

For the missing forces flaw, the focus is on context and frequency. Inter-trial features and background context become tools to prime decision-making similar to how go/no-go trial ratios influence the strength of prepotent motor activity during inhibitory control tasks (Wessel, 2018). The focus is again upon influencing decision-making factors without concern for how one decision affects subsequent decisions.

For the missing contingencies flaw, trained personnel will know the difference between live fire and simulation better than anyone else. The role of consequences may be more illuminating for them given that they have a true understanding of the difference between firing a real weapon versus mimicking an action. One phenomenal missing contingencies argument involves the reliance upon misidentifying harmless objects as a crux of first-person shooter tasks. There are other ways to explore errors in lethal force decisions by intentionally introducing ambiguity into the task (Biggs, Pettijohn, & Gardony, 2021), which shooting paradigms could exploit.

Still, the common missing element across all three flaws is consequence – shooters can fire too slowly without getting hurt or shoot unarmed targets without punishment. Training instructors cannot avoid this topic in the same way as experiments that design around the problem. Handing someone a live weapon versus a plastic toy will inevitably impose some level of stress and anxiety. Rather than avoid the challenge, trainers sometimes address anxiety and realism with non-lethal training ammunition (Taverniers & De Boeck, 2014; Taverniers, Smeets, Van Ruysseveldt, Syroit, & von Grumbkow, 2011). The simple solution is to impose a consequence. Shooters will feel the pain sensation of being shot, and they know their own behavior might inflict pain on someone else (Biggs & Doubrava, 2019). Because the simulation is now conducted against a dynamic and thinking opponent, with the consequence of being shot, the result is a more realistic training environment. It just cannot be easily replicated in a laboratory setting. The challenge is transitioning the experiment to the field conditions rather than trying to make the laboratory more like the field. Find the operational need first, figure out how it is trained, and make the experiment match that scenario. Do not design in reverse and try to find an operational need that fits your experiment without acknowledging the applied limitations of this approach.

By focusing on the operational needs first, and then building a laboratory paradigm to replicate that need, the experimental flaws are far less likely to be overlooked. Methodological issues such as measuring reaction time with training weapons should be overcome with innovation rather than built into studies as experimental flaws. Moreover, the resulting study is more likely to have a real-world consequence as there could be a method to measure results, compare them to existing procedures, and finally integrate changes into training. Begin with a transition plan focused on the end user – and if the experimental flaws are not avoided, they will become clear.

Financial support. The views expressed in this article are those of the author and do not necessarily reflect the official policy or position of the Department of the Navy, Department of Defense, nor the U.S. government. The author is a military service member or employee of the U.S. government. This work was prepared as part of their official duties. Title 17 U.S.C. §105 provides that "Copyright protection under this title is not available for any work of the United States Government." Title 17 U.S.C. §101 defines a U.S. government work as a work prepared by a military service member or employee of the U.S. government as part of that person's official duties.

Conflict of interest. The author has no financial or non-financial competing interests in this manuscript.

References

Biggs, A. T., Cain, M. S., & Mitroff, S. R. (2015). Cognitive training can reduce civilian casualties in a simulated shooting environment. *Psychological Science*, *26*(8), 1164–1176.

Biggs, A., & Doubrava, M. (2019). Superficial ballistic trauma and subjective pain experienced during force-on-force training and the observed recovery pattern. *Military Medicine*, *184*(11–12), e611–e615.

https://doi.org/10.1017/S0140525X21000017 Published online by Cambridge University Press

Biggs, A., Pettijohn, K., & Gardony, A. (2021). When the response does not match the threat: The relationship between threat assessment and behavioural response in ambiguous lethal force decision-making. *Quarterly Journal of Experimental Psychology*, *74*(5), 812–825.
Blacker, K. J., Pettijohn, K. A., Roush, G., & Biggs, A. T. (2021). Measuring lethal force performance in the lab: The effects of simulator realism and participant experience. *Human Factors*, *63*(7), 1141–1155.
Cesario, J., & Carrillo, A. (in press). Racial bias in police officer deadly force decisions: What has social cognition learned? In D. E. Carlston, K. Johnson & K. Hugenberg (Eds.), *The Oxford handbook of social cognition* (2nd ed.). Oxford University Press.
Correll, J., Wittenbrink, B., Park, B., Judd, C. M., & Goyle, A. (2011). Dangerous enough: Moderating racial bias with contextual threat cues. *Journal of Experimental Social Psychology*, *47*, 184–189.
Cox, W. T. L., & Devine, P. G. (2016). Experimental research on shooter bias: Ready (or relevant) for application in the courtroom? *Journal of Applied Research in Memory and Cognition 5*, 236–238.
James, L., Klinger, D., & Vila, B. (2014). Racial and ethnic bias in decisions to shoot seen through a stronger lens: Experimental results from high-fidelity laboratory simulations. *Journal of Experimental Criminology*, *10*, 323–340.
James, L., Vila, B., & Daratha, K. (2013). Results from experimental trials testing participant responses to White, Hispanic and Black suspects in high-fidelity deadly force judgment and decision-making simulations. *Journal of Experimental Criminology 9*, 189–212.
Nieuwenhuys, A., & Oudejans, R. R. (2010). Effects of anxiety on handgun shooting behavior of police officers: A pilot study. Anxiety, Stress, *&* *Coping*, *23*(2), 225–233.
Nieuwenhuys, A., & Oudejans, R. R. (2011). Training with anxiety: Short- and long-term effects on police officers' shooting behavior under pressure. *Cognitive Processing*, *12* (3), 277–288.
Oudejans, R. R. D. (2008). Reality-based practice under pressure improves handgun shooting performance of police officers. *Ergonomics*, *51*(3), 261–273.
Patton, D., & Gamble, K. (2016). Physiological measures of arousal during soldier-relevant tasks performed in a simulated environment. In D. D. Schmorrow & C. M. Fidopiastis (Eds.), *Foundations of augmented cognition: Neuroergonomics and operational neuroscience: 10th international conference*, AC 2016, Held as Part of HCI International 2016, Toronto, ON, Canada, July 17–22, 2016, Proceedings, Part I (pp. 372–382). Springer International Publishing.
Sim, J. J., Correll, J., & Sadler, M. S. (2013). Understanding police and expert performance: When training attenuates (vs. exacerbates) stereotypic bias in the decision to shoot. *Personality and Social Psychology Bulletin*, *39*, 291–304. doi: 10.1177/0146167212473157
Taverniers, J., & De Boeck, P. (2014). Force-on-force handgun practice: An intra-individual exploration of stress effects, biomarker regulation, and behavioral changes. *Human Factors*, *56*(2), 403–413.
Taverniers, J., Smeets, T., Van Ruysseveldt, J., Syroit, J., & von Grumbkow, J. (2011). The risk of being shot at: Stress, cortisol secretion, and their impact on memory and perceived learning during reality-based practice for armed officers. *International Journal of Stress Management*, *18*(2), 113–132.
Taylor, P. L. (2020). Dispatch priming and the police decision to use deadly force. *Police Quarterly*, 1098611119896653.
Wessel, J. R. (2018). Prepotent motor activity and inhibitory control demands in different variants of the go/no-go paradigm. *Psychophysiology*, *55*(3), e12871.

A skeptical reflection: Contextualizing police shooting decisions with skin-tone

David M. Blake

Blake Consulting & Training Group, Brentwood, CA 94513, USA.
Dave@Blake-Consulting.com; https://blake-consulting.com/

doi:10.1017/S0140525X21000613, e70

Abstract

This commentary expands the discussion of Cesario's *Missing Forces Flaw* by identifying and discussing variables that influence police shooting decisions but are often absent from bias-based research. Additionally, the closing identifies novel recommendations for future contextually related research.

Academic research in criminology and social psychology has presented a historical divergence in methods, results, and recommendations in their attempts to operationalize the influence of implicit racial bias on deadly police shootings (Correll, Hudson, Guillermo, & Ma, 2014; Fridell, 2016; Hollis & Jennings, 2018; James, James, & Vila, 2016; Klinger & Slocum, 2017; Nix, Campbell, Byers, & Alpert, 2017 ; Rotello, Kelly, & Heit, 2018; Worrall, Bishopp, & Terrill, 2020). While disparity between studies exists, findings of bias often result in recommendations for organizational reforms such as implicit-bias training or the issuance of body-worn cameras (BWCs). These recommendations persist in the absence of evidence showing an associated reduction in police shootings (Engel, McManus, & Isaza, 2020); Klinger & Slocum, 2017).

Yet, despite academic disagreement on this topic, U.S. law enforcement agencies spend limited training time and economic resources on questionable de-biasing reform efforts (Engel et al., 2020; FitzGerald, Martin, Berner, & Hurst, 2019; Forscher et al., 2019; Klinger & Slocum, 2017; Paluck & Green, 2009). These reform efforts have neither reduced the racial disparity nor the overall number of fatal police shootings (Washington Post Fatal Force Database, 2015–2021). These facts alone should raise the level of skepticism concerning the epistemology of police shooting research when evaluating the influence of racial bias. However, Cesario describes additional flaws found within current research, of which one, the *missing forces flaw*, seems most prudent to expand upon in this narrative.

Cesario identifies and defines the missing forces flaw as a deficiency within implicit bias research. The missing forces flaw, as applied to police shootings, presents as an insufficient inclusion of salient contextual factors that may impact officer's decision to shoot. Cesario mentions some of these contextual factors (i.e., violent crime rates), but many other influences – for example, organizational, supervisory, environmental, and situational – are found in the literature (McFarlane & Amin, 2021). Two important factors influencing police shootings and not mentioned by Cesario are (a) police policy/training and its impact on how a subject's antecedent behavior is perceived, and (b) previous findings from deadly force judgment and decision-making (DFJDM) simulator-based research methods.

Police policy/training and subject antecedent behavior

Police officers across the nation are taught to evaluate the severity of the crime, the level of active resistance from the suspect, and the potential for injury to themselves/others before making a shooting decision (Graham v. Connor, 490 U.S. 386, 1989). Although no known experimental research explicitly examines these variables in aggregate, the last two criteria are well-established as influential to police shooting decisions (Hine, Porter, Westera, Alpert, & Allen, 2019; Shane & Swenson, 2020; Wheeler, Phillips, Worrall, & Bishopp, 2018). In fact, antecedent subject behavior proximal to any police use of force, including shootings, has repeatedly been identified as one of, if not the most, influential factors (Smith, Engel, & Cherkauskas, 2019).

Within the appropriate context, some subject behaviors that influence police shootings include quick or aggressive actions (i.e., furtive movements), closing the distance with officers,

intoxication, being armed or acting as if armed, failing to comply with officer commands, and attacking or attempting to disarm an officer (Aveni, 2008; Fachner & Carter, 2015; Hine et al., 2019; Klinger & Slocum, 2017; Shane & Swenson, 2020). These same subject behaviors are often conceptually associated with shoot/don't-shoot decision-making points found within police law enforcement training (e.g., DFJDM simulators) (James et al., 2016). Therefore, although Cesario cautions against "victim-blaming," the idea of ignoring an individual's antecedent behaviors proximal to a police shooting is, bluntly, nonsensical.

Simulator research

James et al. (2016) improved the ecological validity of studies of racial bias in police shooting decisions using DFJDM simulators. The simulator method allows for a semi-realistic interaction between an officer and a subject using a projected video and laser-based weapons. The simulator method differs from computer-based first-person shooter research use of models (Correll et al., 2014), most notably for its interactive capabilities and replica weapons. Using DFJDM simulators, James and colleagues have consistently found no significant anti-black shooting behaviors by police participants (James et al., 2016; James, James, & Vila, 2018). Hence, the DFJDM simulator research method should create skepticism and drive attempts at replication. Additional reservations are cultivated by evaluating the long list of DFJDM simulator research articles identifying other variables that influence police shooting decisions.

For instance, using the DFJDM simulator method, Nieuwenhuys, Savelsbergh, and Oudejans (2012, 2015) found anxiety induced by a pain-inducing "shoot-back" cannon significantly decreased police shooting response time and increased shooting errors. Other studies using DFJDM simulators found officer experience, subject demeanor, clothing, age, type of crime, and variations in subject movement patterns (i.e., rapid turns) are influential to police shootings (Aveni, 2008; James et al., 2018; Suss & Ward, 2018).

The future: Formalizing the framework

Reality-based police shooting research is arguably a more complicated and therefore an underused method. However, technology and government transparency has provided researchers with a mechanism to conduct naturalistic research. For example, prosecutor's offices across the United States provide shooting memoranda outlining details of a police shooting. Researchers might analyze these documents via qualitative content analysis (QCA) to identify factors influential to a police shooting. Associated BWC footage may also be part of the investigation (Wheeler *et al.*, 2018). A sample of these cases could be analyzed to determine individual variables influencing police shootings. A subsequent comparative analysis between demographic groups that account for these variables may then be evaluated for racial disparity between groups.

In closing, Klinger (2012) reminds us that skepticism of causation is a foundational element of science. The novel framework Cesario suggests - a portion of which I expanded upon in this narrative - provides a foundation for skepticism on the interconnections between implicit racial bias and police shooting decisions. Support for Cesario can be found in the many critical reviews (e.g., methodological flaws) of studies exploring the influence of racial bias on police shootings (Fridell, 2016; Hollis & Jennings, 2018; James et al., 2016; Klinger & Slocum, 2017; Wheeler et al., 2018). Therefore, for both ethical and scientific purposes, researchers should embrace Cesario's narrative not only to better understand group disparity, but also to advance a more rigorous approach to police decision-making research.

Conflict of interest. None.

References

Aveni, T. (2008). A critical analysis of police shootings under ambiguous circumstances. *The MMRMA Deadly Force Project*. Retrieved from: http://www.theppsc.org/Research/V3.MMRMA_Deadly_Force_Project.pdf.

Correll, J., Hudson, S. M., Guillermo, S., & Ma, D. S. (2014). The police officer's dilemma: A decade of research on racial bias in the decision to shoot. *Social and Personality Psychology Compass*, *8*(5), 201–213. doi:10.1111/spc3.12099

Engel, R. S., McManus, H. D., & Isaza, G. T. (2020). Moving beyond "Best practice": Experiences in police reform and a call for evidence to reduce officer-involved shootings. *The Annals of the American Academy of Political and Social Science*, *687*(1), 146–165. https://doi.org/10.1177/0002716219889328.

Fachner, G., & Carter, S. (2015). *An assessment of deadly force in the Philadelphia police department. Collaborative reform initiative.* Office of Community Oriented Policing Services.

FitzGerald, C., Martin, A., Berner, D., & Hurst, S. (2019). Interventions designed to reduce implicit prejudices and implicit stereotypes in real-world contexts: A systematic review. *BMC Psychology*, *7*(1), 1–12. https://doi:10.1186/s40359-019-0299-7

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

Fridell, L. A. (2016). Explaining the disparity in results across studies assessing racial disparity in police use of force: A research note. *American Journal of Criminal Justice*, *42*(3), 502–513. https://doi.org/10.1007/s12103-016-9378-y.

Graham v. Connor, 490 U.S. 386 (1989).

Hine, K. A., Porter, L. E., Westera, N. J., Alpert, G. P., & Allen, A. (2019). What were they thinking? Factors influencing police recruits' decisions about force. *Policing and Society*, *29*(6), 673–691. https://doi:10.21428/1163c5ca.8dfe6a91

Hollis, M. E., & Jennings, W. G. (2018). Racial disparities in police use-of-force: A state-of-the-art review. *Policing: An International Journal*, *41*(2), 178–193. https://doi.org/10.1108/pijpsm-09-2017-0112.

James, L., James, S. M., & Vila, B. J. (2016). The reverse racism effect. *Criminology & Public Policy*, *15*(2), 457–479. https://doi.org/10.1111/1745-9133.12187.

James, L., James, S., & Vila, B. (2018). Testing the impact of citizen characteristics and demeanor on police officer behavior in potentially violent encounters. *Policing: An International Journal of Police Strategies and Management*, *41*(1), 24–40. https://doi.org/10.1108/PIJPSM-11-2016-0159.

Klinger, D. A. (2012). Back to basics: Some thoughts on the importance of organized skepticism in criminology and public policy. *Criminology and Public Policy*. *11*(4), 637–640. doi:10.1111/j.1745-9133.2012.00843.x

Klinger, D. A., & Slocum, L. A. (2017). Critical assessment of an analysis of a journalistic compendium of citizens killed by police gunfire. *Criminology & Public Policy*, *16*(1), 349–362. doi:10.1111/1745-9133.12283

McFarlane, P., & Amin, A. (2021). Investigating fatal police shootings using the human factors analysis and classification framework (HFACS). *Police Practice and Research*, *22*(7), 1777–1791. doi: 10.1080/15614263.2021.1878893.

Nieuwenhuys, A., Savelsbergh, G. J., & Oudejans, R. R. (2012). Shoot or don't shoot? Why police officers are more inclined to shoot when they are anxious. *Emotion*, *12*(4), 827–833. doi:10.1037/a0025699

Nieuwenhuys, A., Savelsbergh, G. J., & Oudejans, R. R. (2015). Persistence of threat-induced errors in police officers' shooting decisions. *Applied Ergonomics*, *48*, 263–272. doi:10.1016/j.apergo.2014.12.006

Nix, J., Campbell, B. A., Byers, E. H., & Alpert, G. P. (2017). A bird's eye view of civilians killed by police in 2015. *Criminology & Public Policy*, *16*(1), 309–340. doi:10.1111/1745-9133.12269

Paluck, E. L., & Green, D. P. (2009). Prejudice reduction: What works? A review and assessment of research and practice. *Annual Review of Psychology*, *60*(1), 339–367. https://doi.org/10.1146/annurev.psych.60.110707.163607.

Rotello, C. M., Kelly, L. J., & Heit, E. (2018). The shape of ROC curves in shooter tasks: Implications for best practices in analysis. *Collabra: Psychology*, *4*(1), 32. https://doi:10.1525/collabra.171

Shane, J., & Swenson, Z. (2020). *Unarmed and dangerous: Patterns of threats by citizens during deadly force encounters with police* (1st ed.). Routledge.

Smith, M., Engel, R.S., & Cherkauskas, J.C. (2019). A multi-method investigation of officer decision-making and force used or avoided in arrest situations: Tulsa, Oklahoma police department administrative data analysis report. Retrieved from

https://www.theiacp.org/sites/default/files/Research%20Center/FINAL%20Tulsa%20Data%20Analysis%20Report-Use%20of%20Force%20Decision-making%20(1).pdf.

Suss, J., & Ward, P. (2018). Revealing perceptual-cognitive expertise in law enforcement: An iterative approach using verbal-report, temporal-occlusion, and option-generation methods. *Cognition, Technology & Work*, *20*(4), 585–596. https://doi.org/10.1007/s10111-018-0493-z.

Wheeler, A. P., Phillips, S. W., Worrall, J. L., & Bishopp, S. A. (2018). What factors influence an officer's decision to shoot? The promise and limitations of using public data. *Justice Research and Policy*, *18*(1), 48–76. https://doi.org/10.1177/1525107118759900.

Worrall, J. L., Bishopp, S. A., & Terrill, W. (2020). The effect of suspect race on police officers' decisions to draw their weapons. *Justice Quarterly*. https://doi:10.1080/07418825.2020.1760331.

Taking social psychology out of context

Michael Brownstein[a], Daniel Kelly[b] and Alex Madva[c]

[a]Department of Philosophy, John Jay College and the Graduate Center, CUNY, New York, NY, USA 10019; [b]Department of Philosophy, Purdue University, West Lafayette, IN, USA 47906-2098 and [c]Department of Philosophy, California State Polytechnic University, Pomona, CA 91768, USA.
msbrownstein@gmail.com, drkelly@purdue.edu, alexmadva@gmail.com
www.michaelsbrownstein.com, http://web.ics.purdue.edu/~drkelly/, https://www.alexmadva.com

doi:10.1017/S0140525X21000704, e71

Abstract

We endorse Cesario's call for more research into the complexities of "real-world" decisions and the comparative power of different causes of group disparities. Unfortunately, these reasonable suggestions are overshadowed by a barrage of non sequiturs, misdirected criticisms of methodology, and unsubstantiated claims about the assumptions and inferences of social psychologists.

We endorse Cesario's call for more research into the complexities of "real-world" decisions and the comparative power of different causes of group disparities (Brownstein, Madva, & Gawronski, 2020; Cesario et al. 2010; Davidson & Kelly, 2020). Unfortunately, these reasonable suggestions are overshadowed by a barrage of non sequiturs, misdirected criticisms of methodology, and unsubstantiated claims about the assumptions and inferences of social psychologists. We leave the latter issue aside, except to express frustration that the purportedly ubiquitous "logic among social psychologists" is documented with a mere three citations (sect. 1., para. 1), while a later discussion of real-world group differences – for example – is supported with twenty-nine (sect. 3.2, para. 2).

Cesario's "Missing Forces Flaw" alleges that social psychologists dismiss potential causes of group disparities other than bias, such as gender differences in science, technology, engineering, and mathematics (STEM) abilities or neighborhood crime rates in the case of police shootings. Far from ignoring such causes, however, many social psychologists *assume* them. A commonplace in social psychology is that biases are symptoms or *mirror-like reflections* of social reality (e.g., Dasgupta, 2013; Forscher et al., 2019; Glaser, 2014; cf. Madva, 2016a, 2017; Payne, Vuletich, & Lundberg, 2017). It makes little sense for Cesario to claim that social psychologists fail to interpret "experimental categorical effects in light of other known forces on group outcomes" (sect. 3.2, para. 4) when social psychologists also argue that experimental categorical effects are reflections of other known forces on group outcomes. We happen to be skeptical of the social determinism implied by talk of "mirror-like reflections," but examining this idea requires more research into the nature of categorical biases and the ways they interact with broader social context, not less.

Acknowledging the need for more research does not, thankfully, commit us to the dubious claim that existing lab studies "cannot" provide information about real-world decisions and group disparities. Cesario's all-or-nothing claims about the in-principle uninformativeness of lab studies obscure more difficult questions about *how much* researchers should update their beliefs about group disparities based on different lab studies. Despite one passing reference to Bayes, Cesario has no discussion of what it can mean for *x* to "provide information about" or "be evidence of" *y*, or, crucially, the difference between deductive, absolutist reasoning and inductive, probabilistic reasoning. Thus, ironically, Cesario inductively infers from one set of limited-information lab studies that other limited-information lab studies are entirely uninformative about the "real world." Instead of accusing social psychologists of drawing fallacious deductive conclusions, perhaps Cesario's criticisms could be reformulated to say that researchers are updating their beliefs sometimes more (when it comes to the explanatory power of bias) and sometimes less (when it comes to the explanatory power of other factors) than they should. But evaluating such claims about more fine-grained epistemic responses to the evolving evidence would require arguments and evidence Cesario hasn't provided.

Cesario also commits a version of the fundamental attribution error he attributes to social psychologists. His view is that lab-based studies on bias ignore wider context. But other than a brief mention of "reward structure" (sect. 8, para. 7), one is left with the impression that social psychologists' fallacious inferences are the cause of the problem. Cesario ignores the myriad structural incentives and constraints – the *context!* – guiding research choices. There is, for example, evidence to suggest that the very-warranted pressure to produce more replicable results has made social psychology less ecologically valid and more reliant on limited-information online studies (Sassenberg & Ditrich, 2019). An alternative version of the target article could have explored the tradeoffs and consequences accompanying these shifting structural incentives.

If correct, Cesario's arguments would impugn not just social psychology, but much of experimental science. In medical and pharmacological research, a decontextualized lab study testing how mice respond to a vaccine provides tentative evidence for how other mammals, like humans, will react outside the lab. Researchers adjust their prior beliefs accordingly, despite much "missing information," and eventually take their research outside the lab. Social psychology lacks something analogous to phase 2 and phase 3 clinical trials presumably because it is not funded by capital or supported by government like medical research, not because of its "logic."

Cesario also accuses social psychologists of "methodological trickery" (sect. 5, para. 5) by treating probabilistic information people use in ordinary life as bias during experiments. But this is not trickery; it isn't even ecologically invalid. There are *many* real-world contexts in which people do and should suspend knowledge of probabilities, for both epistemic and moral reasons

(Madva, 2016b). When serving on a jury, you are reasonably restricted from considering certain information (e.g., the perceived criminality of members of the defendant's social group). Or consider anonymous review in academic journals and "prestige bias." Suppose the prestige of an author's university affiliation predicts, in some way, the quality of her submission. It would still be a separate and legitimate question whether the author's affiliation *should* be taken into consideration by journal editors.

Similarly, it isn't a flaw of an experimental paradigm – or "blank slate worldism" (sect. 5, para. 7) – if it tests whether participants can bracket some of what they know in order to discover something about their minds. Asking participants in a shooter task to ignore background base rates, such as the likelihood, given their race, that a person is holding a gun, is entirely appropriate *for the epistemic aim* of determining that bias exists and for learning how it operates under certain conditions. Learning this about bias is different than learning about what causes it to exist or what effects it has under other conditions, but all of this is worth knowing.

Setting aside the target article's non sequiturs and melodrama, what remains are familiar challenges faced by any science striving to generalize and apply its results. A final irony, then, is that many of the improvements to the experimental and theoretical paradigms that Cesario discusses – simulator studies of shooting decisions, recognition that implicit biases aren't unconscious – are *because* of the kind of work done by social psychologists and their fellow travelers in adjacent disciplines. Continued progress on such challenges will very likely be the result of more, not less, of the relevant research.

Financial support. This research received no specific grant from any funding agency, commercial, or not-for-profit sectors.

Conflict of interest. None.

References

Brownstein, M., Madva, A., & Gawronski, B. (2020). Understanding implicit bias: Putting the criticism into perspective. *Pacific Philosophical Quarterly*, *101*, 276–307.

Cesario, J., Plaks, J. E., Hagiwara, N., Navarrete, C. D., & Higgins, E. T. (2010). The ecology of automaticity: How situational contingencies ahape action semantics and social behavior. *Psychological Science*, *21*(9), 1311–1317. https://doi.org/10.1177/0956797610378685.

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.

Davidson, L., & Kelly, D. (2020). Minding the gap: Bias, soft structures, and the double life of social norms. *Journal of Applied Philosophy, Special Issue on Bias in Context*, *37* (2), 190–210. doi: 10.1111/japp.12351

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

Glaser, J. (2014). *Suspect race: Causes and consequences of racial profiling*. Oxford University Press.

Madva, A. (2016a). A plea for anti-anti-individualism: How oversimple psychology misleads social policy. *Ergo, an Open Access Journal of Philosophy*, *3*(27), 701–728. https://doi.org/10.3998/ergo.12405314.0003.027.

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

Madva, A. (2017). Biased against debiasing: On the role of (institutionally sponsored) self-transformation in the struggle against prejudice. *Ergo, an Open Access Journal of Philosophy*, *4*(6), 145–179. http://dx.doi.org/10.3998/ergo.12405314.0004.006.

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

Sassenberg, K., & Ditrich, L. (2019). Research in social psychology changed between 2011 and 2016: Larger sample sizes, more self-report measures, and more online studies. *Advances in Methods and Practices in Psychological Science*, *2*(2), 107–114. https://doi.org/10.1177/2515245919838781.

Experimental studies of bias: Imperfect but neither useless nor unique

Callie H. Burt[a] and Brian B. Boutwell[b]

[a]Department of Criminal Justice & Criminology, Andrew Young School of Policy Studies, Georgia State University; Center for Research on Interpersonal Violence, Atlanta, GA 30303, USA and [b]Department of Criminal Justice and Legal Studies, University of Mississippi, School of Applied Sciences; Criminal Justice and Legal Studies, University of Mississippi Medical Center, University, MS 38677-1848, USA.
cburt@gsu.edu, bbboutwe@olemiss.edu
www.callieburt.org, https://legalstudies.olemiss.edu/people/brian-boutwell/

doi:10.1017/S0140525X2100087X, e72

Abstract

Cesario provides a compelling critique of the use of experimental social psychology to explain real-world group disparities. We concur with his targeted critique and extend "the problem of missing information" to another common measures of bias. We disagree with Cesario's broader argument that the entire enterprise be abandoned, suggesting instead targeted utilization. Finally, we question whether the critique is appropriately directed at experimental social psychologists.

In his compelling article, Cesario offers a cogent critique of "the widespread use of experimental social psychology to understand real-world group disparities" (abstract). In our reading, Cesario offers both narrow and broad arguments. We concur with the narrow version, which highlights three "fatal flaws" in standard experimental bias studies that undermine their direct contribution to explaining real-world group disparities in social outcomes. This critique does not imply that these studies have no value – we think they do – or that stereotype biases do not exist – of course they do, but rather that experimental evidence of biased associations do not illuminate major causes of group disparities because of a number of limitations clearly outlined in Cesario's article.

Chief among these limitations is what Cesario calls "the problem of missing information." In contrast to these experiments, in the real world, decision-making does not operate in an informational vacuum. The strength of experiments is their control – isolating the effects of one variable by creating an informational vacuum (in this case, only social category membership). Yet these situations – devoid of individual, situational, and contextual information and with time pressures imposed to prevent the activation of conscious processing – are precisely when stereotypes (negative and positive) are relied upon to fill gaps in information. Such stereotypes are influenced not only by media hype and personal experiences, but also, in some cases, knowledge of group average behavioral differences. Thus, the strength of experiments is a weakness when extrapolating to real-world decision-making where stereotypes may not

be activated given the wealth of other contextual information. This, Cesario argues and we agree, not only makes the external validity of the tests questionable, but it renders the null hypothesis of "no difference" potentially unrealistic – or at least requiring justification – given the second critique identified by Cesario – that of missing forces (which we also view as missing information).

Cesario's important critique about missing information is usefully extended to other common methods of measuring biases, including increasingly pervasive self-report discrimination measures, which can suffer from similar limitations, albeit in reverse form. In "experimental task" situations decision-makers only have group membership information. Conversely, in real life, people in interaction differ on many dimensions. In self-report discrimination instruments, individuals are asked to attribute causes of perceived unfair treatment by others usually without knowledge of intent. Individuals may attribute one cause (sex, race, weight, age, etc.) when it may be a different one (appearance, tattoos, and accent), or the perpetrator is grumpy, tired, and treating everyone poorly. This is not to suggest that all discriminatory acts are ambiguous in their source or motivation (e.g., calling someone a racist, sexist, or homophobic slur), only that much captured in self-report discrimination measures is based on attributions without full information. This reliance on perceived intent and attribution distorts measurement to some unknown extent. For example, African American women generally self-report less racial discrimination experiences than Black men (although this varies across specific discrimination type, see Burt & Simons, 2015). This is no doubt due, in part, to ambiguity in attribution of the source of mistreatment. Focusing on sex and race (to make a point), if an African American woman is given much worse service than a white man in front of her, it could be about race, sex, or both, whereas for a Black man, it is about race (see Essed, 1991). We note this limitation both to recognize the pervasive problem of missing information in bias research (and in many domains) as well as the fact that these experimental bias studies are not unique in their missing information problem. Note that this critique does not imply that perceptual measures are useless – indeed, we think they can be quite useful – only that the limitations should be recognized and addressed whenever possible with methodological innovation and triangulation of methods.

While we concur with what we view as the narrow version of Cesario's argument – that these studies do not identify causes of group disparities, we disagree with the broader critique – that the entire enterprise of using experimental social psychology to shed light on group disparities should be abandoned. Rather than abandonment, we suggest targeted utilization where research can use simpler models to identify stereotypes as a starting point for understanding causes of group disparities to be addressed in more comprehensive investigations (e.g., Johnson, Cesario, & Pleskac, 2018). Understanding what stereotypes or implicit biases persist, when these influence attributions and decision-making (e.g., under cognitive load, lack of information, and high threat), and what situational, contextual, and individual information mitigates against the reliance on unconscious stereotypes remains an important research question to which these studies can contribute.

Finally, we note that we are unsure whether the issues Cesario raises are appropriately directed at experimental social psychologists. In our reading, most scholars are cautious in their claims about what these studies can tell us about causes of group disparities. Indeed, the Moss-Racusin, Dovidio, Brescoll, Graham, and Handelsman's (2012) article used as a prototypical example by Cesario was replete with cautionary statements such as "might," "could," "possibly," and with explicit acknowledgement "that various lifestyle choices likely contribute to the gender imbalance in science" (p. 16764), among other caveats. We have seen journalists, activists, and scholars in other domains misrepresent these studies as identifying "major causes" of group disparities, and they would benefit from heeding Cesario's cogent analysis.

In sum, we concur with Cesario's critique about the limits of these experimental studies for identifying major causes of real-world group disparities. We also agree with Cesario that these studies can provide "important information about stereotyping processes." Rather than final answers, we view them as valuable starting points for identifying biases that may influence decisions and, thus, disparities in certain circumstances (limited information, high cognitive load, and time pressures). Follow up work is needed to understand when, where, and how these may influence outcomes, considering full information, contingencies, and behavioral differences. All models are imperfect, and scientists must rigorously and continuously evaluate the validity of models and methods to identify limitations and flaws, making necessary improvements and corrections, especially when findings have social implications.

Financial support. This research received no specific grant from any funding agency, commercial, or not-for-profit sectors.

Conflict of interest. None.

References

Burt, C. H., & Simons, R. L. (2015). Interpersonal racial discrimination, ethnic-racial socialization, and offending: Risk and resilience among African American females. *Justice Quarterly 32*(3):532–570.

Essed, P. (1991). *Understanding everyday racism: An interdisciplinary theory* (Vol. 2). Sage.

Johnson, D. J., Cesario, J., & Pleskac, T. J. (2018). How prior information and police experience impact decisions to shoot. *Journal of Personality and Social Psychology 115* (4):601.

Moss-Racusin, C. A., Dovidio, J. F., Brescoll, V. L., Graham, M. J., & Handelsman, J. (2012). Science faculty's subtle gender biases favor male students. *Proceedings of the National Academy of Sciences 109*(41):16474–16479.

Understanding causal mechanisms in the study of group bias

Dominik Duell[a] and Dimitri Landa[b]

[a]Department of Politics, University of Innsbruck, 6020 Innsbruck, Austria and [b]Department of Politics, New York University, New York, NY 10012, USA.
dominik.duell@ubik.ac.at, dimitri.landa@nyu.edu
www.dominikduell.com, https://wp.nyu.edu/dimitrilanda/

doi:10.1017/S0140525X21000820, e73

Abstract

Causal mechanisms' portability and their predictions in sometimes counterfactual settings point to the value of studies with

details of interactions and/or convenience samples that depart from those in the proximate contexts of the phenomena of interest. The proper role of such contexts must be construed within an explanatory framework attentive to the nature and properties of relevant causal mechanisms.

The analysis of social and psychological mechanisms at the core of complex behavioral phenomena, such as persistence of group disparities (e.g., Duell & Valasek, 2019; Fershtman & Gneezy, 2001; Fryer, Goeree, & Holt, 2005; Haan, Offerman, & Sloof, 2015; Landa & Duell, 2015), is central to contemporary social sciences. Yet some of the important methodological elements of such studies may sometimes appear puzzling. Among such elements are those that concern the differences between features of laboratory studies that seek to instantiate and isolate specific mechanisms and the details of real-world interactions that these studies model. Failure to properly interpret such differences undergirds the following two claims, most recently advanced in Cesario:

(1) That, as a general matter, because laboratory studies draw on subject pools that are different from those in the modeled interactions, the laboratory results cannot effectively speak to real-world contexts with otherwise proximate decision situations.
(2) That (the laboratory) analysis of counterfactual conditions is irrelevant for understanding real-world social facts.

To see the implications of the first claim, suppose, first, that a social mechanism analyzed in the lab gives rise to predictable behavior B following treatment T, and that the experimental analysis shows that introducing treatment T′ leads to change in behavior, to B′. For Claim 1 to have force, the underlying assertion would have to be that *as a rule*, rather than an anomaly, T′ would be just as, if not more, likely to produce, in a sample more proximate to the target context, a change from B to some B′ that is *in the opposite direction* from B relative to B′. As a matter of evidence provided, this assertion is certainly under-determined: While Cesario lays out studies demonstrating that expert shooters tend to show no or little race-based bias in their decision to shoot, this finding is, plainly, not equivalent to an effect in the opposite direction from the seminal shooter bias studies.

Explaining the differences in these studies' findings is important, yet assuming that these differences, let alone putative behavioral patterns with the opposite sign of what is observed in the lab, is the right general expectation is deeply problematic. At the core of the concept of mechanisms is the idea of robust patterns of connections between causes and effects, driven by general properties of psychological, economic, or other social responses. In this way, portability – among others, from the lab to the world outside it; from the context with one set of subjects to a context with another set; and so on – is central to the very concept of mechanisms (Hedström & Ylikoski, 2010; Hitchcock, 2012). In attacking this portability in principle, Claim 1 is, in effect, calling to abandon the study of social mechanisms as such – a position that today should strike many as, at least prima facie, implausible.

Opposing Claim 1 does not take away from the question of why expert shooters are less biased than many other groups. But a better way to conceive of such a question is as a targeted call for resolving a specific anomaly. Laboratory studies, including studies with varieties of convenience samples, frequently establish mechanisms by which group membership relationships inform behavior, and, depending on application and mechanism, sometimes predict in-group bias, sometimes no bias or even over-correction toward out-group favoritism. These predictions form baseline expectations when taken to alternative target populations, but they are, of course, not the final word for understanding behavior within those populations. Further research needs to establish which of the potential mechanisms have greater weight in a particular target population, resolving the anomalies that may arise from the disjunctions of the observed behaviors across populations. In fact, the research on shooters' bias (Correll, Hudson, Guillermo, & Ma 2014; Correll, Park, Judd, & Wittenbrink, 2002; Johnson, Cesario, & Pleskac, 2018) exemplifies just such a practice, moving from a simplified choice situation in a laboratory setting to more and more contextually rich settings to understand why in a particular population, for example the police force, the expressions of a mechanism giving rise to group bias may sometimes depart from the predictions in a convenience sample.

Claim 2 states that the laboratory analysis of counterfactual conditions (e.g., of situations where there are equally violent black and white offenders or equally skilled black and white workers) is irrelevant for understanding patterns of group discrimination. Yet the study of social mechanisms often requires positing counterfactual possibilities as initial conditions that may, by way of the hypothesized mechanism, help explain observables. Just because most companies do not, per Cesario, choose between similarly skilled black and white workers, or most police officers do not consider how to respond to expectationally similar black and white potential offenders, does not mean there would not be bias if such decision situations emerged. The anticipation of such bias and the relevant individuals' responses to that anticipation, including possible underinvestment in productive capacity or in costly compliance with the state, are some of the central building blocks of important social mechanisms that have been posited to help account for the observable patterns of inequality, quite apart from other determinants of asymmetric standing and treatment of different subpopulations (Moro, 2009). We learn whether such mechanisms – engendering what is known as strategic or equilibrium discrimination – are psychologically plausible (Duell & Landa, 2021a) and how they respond to institutional interventions (Duell & Landa, 2021b) by isolating determinants of particular posited mechanisms, while shutting down others that may create confounding effects. Where such determinants include particular existing distributions of attributes within demographic subpopulations, this means positing counterfactual conditions. In this way, a mechanism of strategic discrimination may, for example, be distinguished from that of statistical discrimination – something that would be impossible if experimentalists were to turn away from modeling conditions that are rare or not representative of what one may find contemporaneously outside the lab.

Effectively addressing social ills requires understanding of the causal mechanisms that bring them about, rather than mere descriptions of associations within immediate target populations. Done well, this is, undoubtedly, a painstaking process that demands both theoretical and experimental imagination, but there is little alternative to it.

https://doi.org/10.1017/S0140525X21000017 Published online by Cambridge University Press

Financial support. This research received no specific grant from any funding agency, commercial, or not-for-profit sectors.

Conflict of interest. None.

References

Correll, J., Hudson, S. M., Guillermo, S., & Ma, D. S. (2014). The police officer's dilemma: A decade of research on racial bias in the decision to shoot. *Social and Personality Psychology Compass 8*(5):201–213, https://doi.org/10.1111/spc3.12099.

Correll, J., Park, B., Judd, C. M., & Wittenbrink, B. (2002). The police officer's dilemma: Using ethnicity to disambiguate potentially threatening individuals. *Journal of Personality and Social Psychology 83*(6):1314–1329, https://doi.org/10.1037/0022-3514.83.6.1314.

Duell, D., & Landa, D. (2021a). Strategic discrimination in hierarchies. *The Journal of Politics 83*(2), 560–576. https://doi.org/10.1086/709860.

Duell, D., & Landa, D. (2021b). Alleviating Strategic Discrimination. *Working paper*. https://s18798.pcdn.co/dimitrilanda/wp-content/uploads/sites/7118/2017/07/duellLanda_alleviatingDiscrimination.pdf.

Duell, D., & Valasek, J. (2019). Political polarization and selection in representative democracies. *Journal of Economic Behavior & Organization 168*:132–165, https://doi.org/10.1016/j.jebo.2019.10.004.

Fershtman, C., & Gneezy, U. (2001). Discrimination in a segmented society: An experimental approach. *The Quarterly Journal of Economics 116*(1):351–377, https://doi.org/10.1162/003355301556338.

Fryer, R. G., Goeree, J. K., & Holt, C. A. (2005). Experience-based discrimination: Classroom games. *The Journal of Economic Education 36*(2):160–170, https://doi.org/10.3200/JECE.36.2.160-170.

Haan, T., Offerman, T., & Sloof, R. (2015). Discrimination in the labour market: The curse of competition between workers. *The Economic Journal 127*(603):1433–1466, https://doi.org/10.1111/ecoj.12352.

Hedström, P., & Ylikoski, P. (2010). Causal mechanisms in the social sciences. *Annual Review of Sociology 36*(1):49–67, https://doi.org/10.1146/annurev.soc.0128https://doi.org/10.1146/annurev.soc.012809.10263209.102632.

Hitchcock, C. (2012). Portable causal dependence: A tale of consilience. *Philosophy of Science 79*(5):942–951, https://doi.org/10.1086/667899.

Johnson, D. J., Cesario, J., & Pleskac, T. J. (2018). How prior information and police experience impact decisions to shoot. *Journal of Personality and Social Psychology 115*(4):601, https://doi.org/10.1037/pspa0000130.

Landa, D., & Duell, D. (2015). Social identity and electoral accountability. *American Journal of Political Science 59*(3):671–689, https://doi.org/10.1111/ajps.12128.

Moro, A. (2009). Statistical discrimination. In: *The new Palgrave dictionary of economics*, eds. S. N. Durlauf and L. E. Blume, Palgrave Macmillan. https://doi.org/10.1057/978-1-349-95121-5.

Beyond stereotypes: Prejudice as an important missing force explaining group disparities

Iniobong Essien[a,b], Marleen Stelter[a,c], Anette Rohmann[a] and Juliane Degner[c]

[a]Department of Psychology, FernUniversität in Hagen, 58097 Hagen, Germany; [b]Department of Social and Organisational Psychology of Social Work, Leuphana Universität Lüneburg, 21335 Lüneburg, Germany and [c]Department of Social Psychology, Universität Hamburg, 20146 Hamburg, Germany.
iniobong.essien@leuphana.de
https://www.leuphana.de/en/institutes/ifsp/team/iniobong-essien.html
marleen.stelter@fernuni-hagen.de
https://www.fernuni-hagen.de/psychologische-methodenlehre/team/marleen.stelter.shtml
anette.rohmann@fernuni-hagen.de
https://www.fernuni-hagen.de/community-psychology/team/
juliane.degner@uni-hamburg.de
https://www.psy.uni-hamburg.de/arbeitsbereiche/sozialpsychologie

doi:10.1017/S0140525X21000832, e74

Abstract

We comment on Cesario's assertion that social psychological intergroup research focuses solely on *stereotypes*, neglecting actual differences between groups to explain group disparities. This reasoning, however, misses yet another explaining force: In addition to stereotypes, ample laboratory and field research documents relationships between group disparities, discrimination, and *prejudice*, which cannot be explained by people's accurate judgments of real-world group differences.

Cesario's analysis of a *Missing Forces Flaw* in experimental research implies that social psychology equates intergroup bias with group stereotypes and conceptualizes stereotypes as the *sole* factor underlying group disparities (e.g., regarding policing outcomes; science, technology, engineering, and mathematics (STEM) participation; and school discipline). This analysis lacks consideration of many critical elements from the intergroup literature. Ample social psychological theorizing and research suggests that discrimination and the resulting group disparities are not only related to stereotypes (i.e., representations of characteristics of social groups), but also to various forms of *prejudice* (e.g., Dixon, Levine, Reicher, & Durrheim, 2012), conceptualized as evaluative, affective, or emotional responses to social groups (Fazio, Jackson, Dunton, & Williams, 1995; Gawronski & Bodenhausen, 2006), ingroup favoritism (Brewer, 1999; Greenwald & Pettigrew, 2014), or dehumanization of outgroups (e.g., Haslam & Loughnan, 2014).

Cesario's oversimplified depiction of social psychology ignores the tripartite attitude framework (e.g., Eagly & Chaiken, 1993), which has been dominant in intergroup research and theorizing since the 1990s (e.g., Haddock, Zanna, & Esses, 1993; Jackson et al., 1996). According to this framework, intergroup attitudes contain cognitive, affective, and behavioral components, typically conceptualized as stereotypes, prejudice, and discrimination. Studies have repeatedly documented that these components account for unique variance in group attitudes (e.g., Haddock et al., 1993; Stangor, Sullivan, & Ford, 1991). Most importantly, prejudice does not merely follow from stereotypes. For example, recent experimental studies support bidirectional causal relations between prejudice and stereotypes (e.g., Kurdi, Mann, Charlesworth, & Banaji, 2019a; Phills, Hahn, & Gawronski, 2020). Some theoretical accounts even presume that prejudice and stereotypes are unrelated because they arise from fundamentally distinct semantic versus affective processes (e.g., Amodio & Devine, 2006; Brigham, 1971). Consequently, prejudice and stereotypes have been conceptualized as both *antecedents* and *consequences* of discriminative behaviors and group disparities.

The target article's depiction of experimental social psychology does not capture these theoretical complexities nor does it consider prejudice as an important missing force explaining group disparities. Experimental research has provided ample evidence that prejudice relates to discriminatory intergroup behaviors. One recent meta-analysis found that racial prejudice was related to discriminatory workplace outcomes (e.g., regarding selection and performance evaluation; Jones et al., 2017). Another meta-analysis even concluded that racial prejudice tends to be "twice as closely" related to discrimination than stereotypes or beliefs (Talaska, Fiske, & Chaiken, 2008, p. 263). Furthermore, meta-analyses on experimental studies on implicit cognition and micro-level interracial interactions suggest that prejudice is related to

subtle behavioral effects, although average effects vary substantially (e.g., Cameron, Brown-Iannuzzi, & Payne, 2012; Greenwald, Poehlman, Uhlmann, & Banaji, 2009; Kurdi et al., 2019b; Oswald, Mitchell, Blanton, Jaccard, & Tetlock, 2013).

Of course, lab research cannot provide all answers regarding real-world group disparities. Consequently, recent empirical approaches have begun focusing on macro-level relationships between intergroup bias and real-world group disparities, consistent with the idea that "racism is more than the sum of the prejudice held by individuals in a system" (Wessells & Dawes, 2006, p. 271). Many of these studies are inspired by the bias of crowds model (Payne, Vuletich, & Lundberg, 2017) and rely on a novel approach, in which individual measures of stereotypes and prejudice are aggregated at geographic levels (e.g., U.S. counties) to investigate their associations with societal outcomes. For example, Riddle and Sinclair (2019) demonstrated that racial disparities in school disciplinary outcomes were related to regional-level prejudice: Black students were more likely disciplined in U.S. counties with higher levels of racial prejudice by White residents, and this effect was robust across a number of metrics of school discipline. Using a similar approach with massive datasets of over 100 million police traffic stops, Stelter, Essien, Sander, and Degner (2022) observed that Black drivers were disproportionately stopped in U.S. counties with higher levels of racial prejudice and threat stereotypes. These relationships were stronger and more robust for measures of prejudice than for measures of threat stereotypes. Furthermore, Hehman et al. (2018) observed that Black people were disproportionately killed by police in regions with higher levels of stereotyping and (to a lesser extent) prejudice by Whites. Importantly, these relationships were even observed when controlling for local violent crime rates. Lastly, macro-level studies have observed relationships between self-reported prejudice and racial disparities in health outcomes (e.g., regarding circulatory diseases; preterm births; Leitner, Hehman, Ayduk, & Mendoza-Denton, 2016; Orchard & Price, 2017). Such findings contradict assumptions about real-world group differences and stereotype accuracy as a major missing force explaining racial disparities in school disciplinary policy, policing, and other societal outcomes.

Together, findings from both micro- and macro-level studies suggest that prejudice is an important force explaining discriminatory behavior, potentially affecting group disparities. These findings have important implications, because they demonstrate that discrimination is not only related to how people *think* about stigmatized groups (i.e., stereotypes), but also to how people *feel* about stigmatized groups (i.e., prejudice). For this reason, we disagree with the target article's assessment that "the information that [people] … have come to learn as being probabilistically accurate in their daily lives" (sect. 5, para. 4) should be regarded as the major missing force explaining group disparities in the lab or field.

In conclusion, Cesario is correct to point out limitations to the interpretability and external validity of experimental social psychological research, and we agree with the target article's assessment that real-world phenomena necessitate multi-causal explanations. But we do not see the call to *abandon* experimental research about group disparities as justified. Instead, a systematic combination of experimental research and field studies should enhance the ecological validity of social psychology research (Dasgupta & Stout, 2012; Mortensen & Cialdini, 2010) and investigate relationships between stereotype- or prejudice-related behavior and group disparities. Ideally, field observations of real-world phenomena are supplemented with additional information (e.g., by decision makers), whereas experimental research on basic mechanisms of intergroup processes might benefit from linking it more closely to behavioral contingencies observed in the real world. Such a full-cycle integration of experimental and field research would be best positioned to further our understanding of the causes of real-world group disparities and help develop effective interventions to reduce them.

Financial support. This research received no specific grant from any funding agency, commercial, or not-for-profit sectors.

Conflict of interest. None.

References

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, https://doi.org/10.1037/0022-3514.91.4.652.

Brewer, M. B. (1999). The psychology of prejudice: Ingroup love and outgroup hate? *Journal of Social Issues 55*:429–444, https://doi.org/10.1111/0022-4537.00126.

Brigham, J. C. (1971). Ethnic stereotypes. *Psychological Bulletin 76*:15–38, https://doi.org/10.1037/h0031446.

Cameron, C., Brown-Iannuzzi, J. L., & Payne, B. (2012). Sequential priming measures of implicit social cognition: A meta-analysis of associations with behavior and explicit attitudes. *Personality and Social Psychology Review 16*:330–350, https://doi.org/10.1177/1088868312440047.

Dasgupta, N., & Stout, J. G. (2012). Contemporary discrimination in the lab and field: Benefits and obstacles of full-cycle social psychology. *Journal of Social Issues 68*:399–412, https://doi.org/10.1111/j.1540-4560.2012.01754.x.

Dixon, J., Levine, M., Reicher, S., & Durrheim, K. (2012). Beyond prejudice: Are negative evaluations the problem and is getting us to like one another more the solution? *Behavioral and Brain Sciences 35*:411–425, https://doi.org/10.1017/S0140525X12001550.

Eagly, A. H., & Chaiken, S. (1993). *The psychology of attitudes*. Harcourt Brace Jovanovich College Publishers.

Fazio, R. H., Jackson, J. R., Dunton, B. C., & Williams, C. J. (1995). Variability in automatic activation as an unobstrusive measure of racial attitudes: A bona fide pipeline? *Journal of Personality and Social Psychology 69*:1013–1027, https://doi.org/10.1037/0022-3514.69.6.1013.

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, https://doi.org/10.1037/0033-2909.132.5.692.

Greenwald, A. G., & Pettigrew, T. F. (2014). With malice toward none and charity for some: Ingroup favoritism enables discrimination. *American Psychologist 69*:669–684, https://doi.org/10.1037/a0036056.

Greenwald, A. G., Poehlman, T. A., Uhlmann, E. L., & 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.http://dx.doi.org/10.1037/a0015575.

Haddock, G., Zanna, M. P., & Esses, V. M. (1993). Assessing the structure of prejudicial attitudes: The case of attitudes toward homosexuals. *Journal of Personality and Social Psychology 65*:1105–1118, https://doi.org/10.1037/0022-3514.65.6.1105.

Haslam, N., & Loughnan, S. (2014). Dehumanization and infrahumanization. *Annual Review of Psychology 65*:399–423, https://doi.org/10.1146/annurev-psych-010213-115045.

Hehman, E., Flake, J. K., & Calanchini, J. (2018). Disproportionate use of lethal force in policing is associated with regional racial biases of residents. *Social Psychological and Personality Science 9*:393–401, https://doi.org/10.1177/1948550617711229.

Jackson, L. A., Hodge, C. N., Gerard, D. A., Ingram, J. M., Ervin, K. S., & Sheppard, L. A. (1996). Cognition, affect, and behavior in the prediction of group attitudes. *Personality and Social Psychology Bulletin 22*:306–316, https://doi.org/10.1177/0146167296223009.

Kurdi, B., Mann, T. C., Charlesworth, T. E. S., & Banaji, M. R. (2019a). The relationship between implicit intergroup attitudes and beliefs. *Proceedings of the National Academy of Sciences of the United States of America 116*:5862–5871, https://doi.org/10.1073/pnas.1820240116.

Jones, K. P., Sabat, I. E., King, E. B., Ahmad, A., McCausland, T. C., & Chen, T. (2017). Isms and schisms: A meta-analysis of the prejudice-discrimination relationship across racism, sexism, and ageism. *Journal of Organizational Behavior 38*(7), 1076–1110. https://doi.org/10.1002/job.2187.

Kurdi, B., Seitchik, A. E., Axt, J., Carroll, T., Karapetyan, A., Kaushik, N., … Banaji, M. R. (2019b). Relationship between the implicit association test and intergroup behavior: A meta-analysis. *American Psychologist 74*:569–586, https://doi.org/10.1037/amp0000364.

Leitner, J. B., Hehman, E., Ayduk, O., & Mendoza-Denton, R. (2016). Blacks' death rate due to circulatory diseases is positively related to Whites' explicit racial bias: A nationwide investigation using project implicit. *Psychological Science 27*:1299–1311, https://doi.org/10.1177/0956797616658450.

Mortensen, C. R., & Cialdini, R. B. (2010). Full-cycle social psychology for theory and application. *Social and Personality Psychology Compass 4*:53–63, https://doi.org/10.1111/j.1751-9004.2009.00239.x.
Orchard, J., & Price, J. (2017). County-level racial prejudice and the black-white gap in infant health outcomes. *Social Science & Medicine 181*:191–198, https://doi.org/10.1016/j.socscimed.2017.03.036.
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, https://doi.org/10.1037/a0032734.
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, https://doi.org/10.1080/1047840X.2017.1335568.
Phills, C. E., Hahn, A., & Gawronski, B. (2020). The bidirectional causal relation between implicit stereotypes and implicit prejudice. *Personality and Social Psychology Bulletin 46*:1318–1330, https://doi.org/10.1177/0146167219899234.
Riddle, T., & Sinclair, S. (2019). Racial disparities in school-based disciplinary actions are associated with county-level rates of racial bias. *Proceedings of the National Academy of Sciences of the United States of America 116*:8255–8260, https://doi.org/10.1073/pnas.1808307116.
Stangor, C., Sullivan, L. A., & Ford, T. E. (1991). Affective and cognitive determinants of prejudice. *Social Cognition 9*:359–380, https://doi.org/10.1521/soco.1991.9.4.359.
Stelter, M., Essien, I., Sander, C., & Degner, J. (2022). Racial bias in police traffic stops: White residents' county-level prejudice and stereotypes are related to disproportionate stopping of Black drivers. *Psychological Science*, *33*(4), 483–496, https://doi.org/10.1177/09567976211051272.
Talaska, C. A., Fiske, S. T., & Chaiken, S. (2008). Legitimating racial discrimination: Emotions, not beliefs, best predict discrimination in a meta-analysis. *Social Justice Research 21*:263–396, https://doi.org/10.1007/s11211-008-0071-2.
Wessells, M., & Dawes, A. (2006). Macro-level interventions: Psychology, social policy, and societal influence processes. In: *Toward a global psychology* eds. M. J. Stevens & U. P. Gielen, pp. 267–298. Psychology Press.

Accuracy in social judgment does not exclude the potential for bias

Jonathan B. Freeman[a], Kerri L. Johnson[b] and Steven J. Stroessner[b]

[a]Department of Psychology, Columbia University, New York, NY 10027, USA and [b]Department of Communication, University of California, Los Angeles, Los Angeles, CA 90095, USA.
jon.freeman@columbia.edu

doi:10.1017/S0140525X2100073X, e75

Abstract

Cesario claims that all bias research tells us is that people "end up using the information they have come to learn as being probabilistically accurate in their daily lives" (sect. 5, para. 4). We expose Cesario's flawed assumptions about the relationship between accuracy and bias. Through statistical simulations and empirical work, we show that even probabilistically accurate responses are regularly accompanied by bias.

We applaud Cesario's appeal to increase the realism of social psychological science and his plea for greater appreciation of effect sizes. However, Cesario's more fundamental critiques of social psychology's research on group bias hinge on misguided theoretical assumptions and fundamental errors. Cesario describes a "Standard Paradigm" in bias research that, he argues, suffers from three flaws. While we take issue with each of these arguments, we focus here on his "Flaw of Missing Forces" (sect. 3 and Table 1) – perhaps the most controversial of the three.

First, Cesario misrepresents the research he describes. Contrary to Cesario's claims, few studies explicitly explore the link between implicit bias and real-world group disparities. Instead, most bias research aims to document group-based distinctions in individuals' decisions, over and above whatever disparities exist in the real world. For example, it is valuable to know whether individuals use gender as a heuristic in science, technology, engineering, and mathematics (STEM) admissions and hiring decisions because demonstrating such a bias illuminates one factor contributing to gender-based differences in STEM representation. Cesario creates a strawman by suggesting that bias research has failed to offer single-factor explanations for complex phenomena. In our view, that is rarely, if ever, the goal of bias research.

Cesario makes a more egregious error by implying that any accuracy in decision-making obviates bias or the need to study it. He argues that bias researchers ignore "the behavior of the targets themselves and the cognitive, motivational, and behavioral differences that exist across groups" (sect. 3.2, para. 1). He concludes that bias research merely tells us that "people learn the conditional probabilities of the behavior of different groups" (sect. 5, para. 3) and what is "probabilistically accurate in their daily lives" (sect. 5, para. 4). Thus, Cesario claims that group-based distinctions in decision-making are an accurate and rational response to social reality. His analysis implies a zero-sum tradeoff between accuracy and bias. We challenge these assertions on both empirical and fundamental statistical grounds.

Existing evidence shows that accuracy is regularly accompanied by bias and, furthermore, that even "probabilistically accurate" responses allow significant opportunity for error-prone behavior. For example, although there is considerable variability in the physical attributes of gay men and lesbians, evidence shows that members of these groups, on average, appear more gender-atypical than their heterosexual counterparts. Moreover, perceivers stereotypically assume gay men and lesbians possess gender-atypical attributes and use these stereotypes to judge others' sexual orientation. Such judgments, according to Cesario, could be construed as a rational response to social reality, negating the need to identify bias in these judgments. However, research shows that using such stereotypes increases accuracy while simultaneously producing bias and overgeneralization (Freeman, Johnson, Ambady, & Rule, 2010; Johnson, Gill, Reichman, & Tassinary, 2007; Stern, West, Jost, & Rule, 2013). When judging targets who do not conform to stereotypes, participants predictably misapply these stereotypes and make erroneous judgments (Freeman et al., 2010). Similar effects have been observed in other forms of visually based social judgment (e.g., Carpinella & Johnson, 2013; Rule, Garrett, & Ambady, 2010). Of course, this is hardly a new idea: Tversky and Kahneman (1974, p. 1131) noted long ago that heuristics such as stereotypes are "highly economical and usually effective, but they lead to systematic and predictable errors." Moreover, the existence of probabilistically accurate responses accompanied by predictable errors is reflected in classic Brunswikian theory and conventional models of human judgment (Hogarth & Karelaia, 2007). Thus, while some stereotypes can result in more accurate responses in the aggregate, they can also increase systematic biases that warrant scrutiny.

We leveraged probability theory in the context of Cesario's centerpiece example of racial bias in the first-person-shooter-task (FPST) to illuminate these patterns. Across 45 simulated FPST experiments, we impose the controversial group differences Cesario describes: that Black people are more armed than White people in the real world (Fig. 1). Our simulations show that, while decision-makers' use of such "real-world" statistics does increase

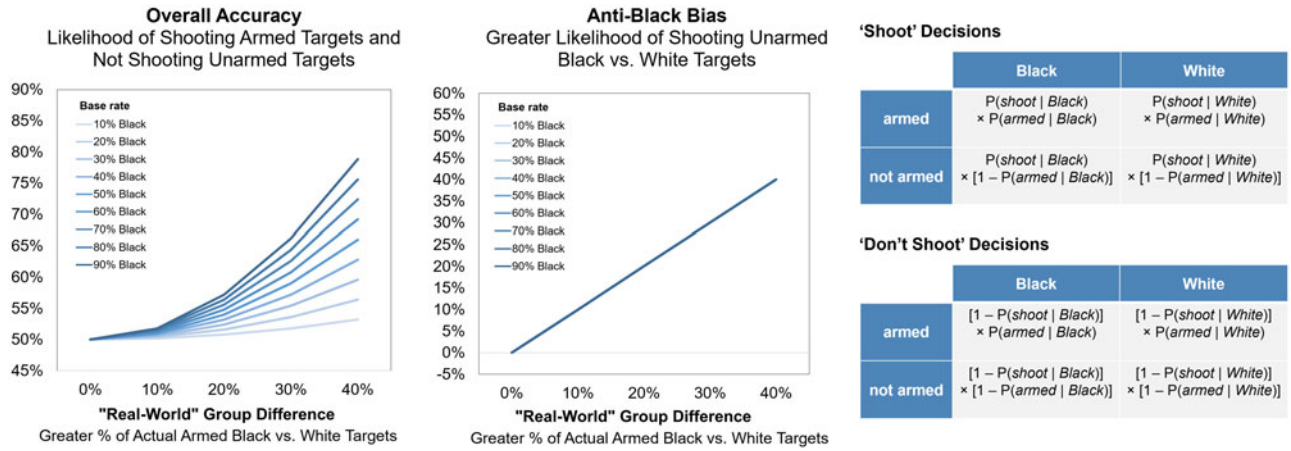

Figure 1 (Freeman et al.). We varied the probability of Black people being armed, *P*(*armed*|*Black*), 50–90%, with *P*(*armed*|*White*) fixed at 50%. Given Cesario's claims about base rates in police encounters, we also varied *P*(*Black*) 10–90%. Per Cesario, we assume that participants accurately encode "real-world" statistics; thus, participants decide to shoot targets based on the likelihood that a target's racial group is armed in the environment: *P*(*shoot*|*Black*) = *P*(*armed*|*Black*) and *P*(*shoot*|*White*) = *P*(*armed*|*White*). Per Cesario, we have reproduced these conditional probabilities in the experimental context. Thus, if Black people are armed at a rate of 70% in the "real-world," which participants encode, then 70% of Black targets in the experiment are armed. As the group difference [*P*(*armed*|*Black*) > *P*(*armed*|*White*)] grew larger, overall accuracy increased, but so did anti-Black bias. A higher base rate (proportion of Black relative to White trials in the experiment) intensified these increases in overall accuracy but did not influence anti-Black bias. Thus, with larger group differences that are accurately encoded, overall accuracy increases, but so does bias.

overall accuracy (i.e., likelihood of shooting only people who are armed), it also increases the rate of racial bias (i.e., greater likelihood of shooting unarmed targets when Black rather than White). Note that this general pattern would be observed even if diagnostic visual cues (e.g., weapon) were permitted to play a role as well; so long as race information is used, accuracy and bias are linked. Thus, if real-world group differences exist, encoding them can improve general accuracy, as Cesario implies, but it cannot eliminate bias. Cesario suggests that investigating bias when people are generally accurate is unnecessary. Quite the opposite, we argue that probabilistically accurate responses are regularly accompanied by predictable errors and overgeneralized stereotyping.

Cesario is incorrect in arguing that target-driven differences between groups are a "missing force" that invalidates decision-makers' bias or the need to study it. Using past empirical work and basic probability theory, we have shown that, even if group differences exist and people take note of them, that knowledge will regularly be misapplied and result in bias. Thus, understanding how flawed individual decision-making plays a role in disparate group outcomes is a worthwhile endeavor. Whatever additional forces create real-world group disparities, people have the opportunity to amplify or attenuate those disparities through their judgment and behavior.

Financial support. This work was supported by National Science Foundation grants BCS-1654731 to J.B.F. and BCS-2017245 to K.L.J.

Conflict of interest. None.

References

Carpinella, C. M., & Johnson, K. L. (2013). Appearance-based politics: Sex-typed facial cues communicate political party affiliation. *Journal of Experimental Social Psychology*, *49*(1), 156–160.

Freeman, J. B., Johnson, K. L., Ambady, N., & Rule, N. O. (2010). Sexual orientation perception involves gendered facial cues. *Personality and Social Psychology Bulletin*, *36*, 1318–1331.

Hogarth, R. M., & Karelaia, N. (2007). Heuristic and linear models of judgment: Matching rules and environments. *Psychological Review*, *114*(3), 733.

Johnson, K. L., Gill, S., Reichman, V., & Tassinary, L. G. (2007). Swagger, sway, and sexuality: Judging sexual orientation from body motion and morphology. *Journal of Personality and Social Psychology*, *93*(3), 321–334.

Rule, N. O., Garrett, J. V., & Ambady, N. (2010). On the perception of religious group membership from faces. *PLoS One*, *5*(12), e14241.

Stern, C., West, T. V., Jost, J. T., & Rule, N. O. (2013). The politics of Gaydar: Ideological differences in the use of gendered cues in categorizing sexual orientation. *Journal of Personality and Social Psychology*, *104*(3), 520.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124–1131.

Centering the relationship between structural racism and individual bias

Agustín Fuentes[a], Laurence Ralph[a] and Dorothy E. Roberts[b]

[a]Department of Anthropology, Princeton University, Princeton, NJ 08540, USA and [b]Carey Law School, University of Pennsylvania, Philadelphia, PA 19104, USA.

afuentes2@princeton.edu (primary contact), lralph@princeton.edu, dorothyroberts@law.upenn.edu
https://anthropology.princeton.edu/people/faculty/agustin-fuentes, https://laurenceralphauthor.com, https://www.law.upenn.edu/cf/faculty/roberts1/

doi:10.1017/S0140525X21000698, e76

Abstract

Cesario misrepresents or ignores data on real-world racist and sexist patterns and processes in an attempt to discredit the assumptions of implicit bias experimentation. His position stands in stark contradiction to substantive research across the social sciences recognizing the widespread, systematic, and structuring processes of racism and sexism. We argue for centering the relationship between structural racism and individual bias.

There is racial and sex/gender inequity in the United States (Grusky, 2019; Healey, Stepnick, & Eileen, 2008). Social

psychologists assert implicit bias plays a role in creating these inequities. Cesario contends one cannot draw substantive conclusions about real-world action from implicit bias research, arguing proponents must show that implicit bias, in and of itself, can be directly responsible for racial and sex/gender inequities – a straw-man argument. Implicit bias is not the only, or the dominant, factor generating discriminatory processes and outcomes of systemic racism and sexism in the United States. But it plays a role.

The purpose of our critique is not to defend implicit bias research or testing. Rather, we contest Cesario's obfuscation of racist and sexist realities in the United States. Cesario misrepresents or ignores data on real-world racism and sexism in an attempt to discredit the assumptions of implicit bias experimentation. His position stands in stark contradiction to decades of research demonstrating how social constructs in the forms of concepts, beliefs, and stereotypes shape action in the context of racism and sexism (Rosa & Díaz, 2020; Krieger, 2020; Roberts & Rollins, 2020).

Cesario argues that implicit bias tests are invalid because the "subjects" might justifiably elicit biased reactions. He asserts "the behavior of the targets themselves and the cognitive, motivational, and behavioral differences that exist across groups" (sect. 3.2, para. 1) may be what is most salient. He argues that not including these "real" differences between races and sex/genders "may leave experimental participants with no useful information to render a judgment other than the target's social category." Cesario offers fatal police encounters as an example, ignoring extensive empirical research demonstrating real-world disproportionate violence against and discrimination toward Black individuals by police. He seeks to devalue the reality of these data, suggesting it is the neighborhood context and the behavior of Black individuals that play the dominant role in explaining why Black people are disproportionately shot and killed by police. Cesario states that being more violent and having higher rates of exposure to police by engaging in more violent crime "greatly – if not entirely – accounts for the overall per capita disparities in being fatally shot by police." This is not accurate in regard to overall crime patterns (FBI UCR, n.d.) or in the fact that police bias against Black individuals is marked and repeatedly demonstrated across multiple geographies, social, economic, and otherwise (Dunham & Petersen, 2017; Hehman, Flake, & Calanchini, 2018; Swencionis & Goff, 2017). Cesario fails to take into account that the reason Black individuals encounter police at higher rates is largely because police departments target segregated Black neighborhoods for greater surveillance and intervention (Gordon, 2020). Police violence is *structured* to impact Black individuals more.

Officers' justifications for shooting to kill an unarmed Black person are often based on a racially biased judgment. Officers defend their lethal actions by claiming they "feared for their life," often using racial stereotypes to convince a jury that it was "reasonable" to feel afraid. For example, white officer Darren Wilson persuaded a St. Louis County grand jury not to indict him for murder after killing 18-year-old Michael Brown Jr. in part by testifying that Brown "looked like a demon" (Waldman, 2014). The U.S. Supreme Court's 1985 decision, *Tennessee v. Garner*, held that police may use deadly force to prevent the escape of a fleeing suspect if the officer has a good-faith belief that the suspect poses a significant threat of death or serious physical injury to the police officer or others. In the 4 years preceding the decision, "officer under attack" was cited in just 33% of police killings; 20 years later, it was cited 62% of the time, becoming an almost infallible means for police officers to defend themselves (Ralph, 2019).

Cesario does note "demographic groups in the U.S. continued to obtain unequal outcomes," but states that this is surprising as there is "little overt, official discrimination for several decades (and in places like academia, preferential policies in favor of underrepresented groups), coupled with increasingly egalitarian attitudes." Cesario's assertion is belied by research demonstrating significant bias and discrimination along racial and sex/gender categories in the academy and across multiple professional contexts without there necessarily being overt/official racism or sexism (Pager & Shepherd, 2008; Small & Pager, 2020). Cesario's misleading assertion about the "real world" presents a view of the "reality" of racial and sex/gender discrimination in the United States that stretches beyond the structural and intellectual merits of implicit bias tests.

Cesario attempts to avoid the critique we offer by stating that he is *only* asking the question of whether decision-maker bias produces group disparities in the immediate outcomes of that decision. He acknowledges that "decision-maker bias may enter earlier in the chain of events" – that racism and sexism may enter in at some point, but that decision maker bias in the implicit bias test is not reflective of, or connected to, such processes. Such a position creates an artificial line dividing individual "in the moment" and systemic processes that serves his argument but does not reflect real-world processes. Serious discussion on this topic must engage the scholarship demonstrating systemic bias affects decision-making (e.g., Amutah et al., 2021; Bailey et al., 2021; Beliso De Jesús, 2020; Dror et al., 2021; Schlosser, 2013). There may be conceptual problems and methodological weaknesses with implicit bias tests, but it is another thing to argue that implicit bias tests in isolation must explain real-world discriminatory actions and outcomes or they are invalid.

Cesario concludes that his goal is to "correct some of the misleading claims about the human mind that have extended out from academia in the last two decades." Here is the true rationale for the article – to challenge the recent convergence across the social sciences of recognition of the widespread, systematic, and structuring processes of racism and sexism. A compelling body of research demonstrates that police often act on their racial biases and justify it on racially biased grounds. Rather than deny this reality, we should explore how implicit bias research – and its critics – can center the relationship between structural racism and individual bias.

Financial support. We received no funding for the preparation and writing of this commentary.

Conflict of interest. We have no conflicts of interest.

References

Amutah, C., Greenidge, K., Mante, A., Munyikwa, M., Surya, S. L., Higginbotham, E. … Aysola, S. (2021). Misrepresenting race – The role of medical schools in propagating physician bias. *New England Journal of Medicine*, *384*, 872–878. doi: 10.1056/NEJMms2025768.

Bailey, Z. D., Feldman, J. M., & Bassett, M. T. (2021). How structural racism works – Racist policies as a root cause of U.S. Racial health inequities. *New England Journal of Medicine*, *384*, 768–773. doi: 10.1056/NEJMms2025396.

Beliso De Jesús, A. M. (2020). The jungle academy: Molding white supremacy in American police recruits. *American Anthropologist*, *122*(1), 143–156.

Dror, I., Melinek, J., Arden, J. L., Kukucka, J., Hawkins, S., Carter, J., & Atherton, D. S. (2021). Cognitive bias in forensic pathology decisions. *Journal of Forensic Science*, *66* (5), 1751–1757. https://doi.org/10.1111/1556-4029.14697.

https://doi.org/10.1017/S0140525X21000017 Published online by Cambridge University Press

Dunham, R. G., & Petersen, N. (2017). Making black lives matter: Evidence-based policies for reducing police bias in the use of deadly force. *Criminology & Public Policy*, *16*, 341.
FBI UCR. (n.d.). *2019 Crime in the United States*. https://ucr.fbi.gov/crime-in-the-u.s/2019/crime-in-the-u.s.-2019/tables/table-43.
Gordon, D. (2020). The police as place-consolidators: The organizational amplification of urban inequality. *Law & Social Inquiry*, *45*(1), 1–27. doi: 10.1017/lsi.2019.31
Grusky, D. (2019). *Social stratification, class, race, and gender in sociological perspective*. Routledge.
Healey, J. F., Stepnick, A., & Eileen, O. (2008). *Race, ethnicity, gender, and class: The sociology of group conflict and change*. Sage Publications.
Hehman, E., Flake, J. K., & Calanchini, J. (2018). Disproportionate use of lethal force in policing is associated with regional racial biases of residents. *Social Psychological and Personality Science*, *9*(4), 393–401.
Krieger, N. (2020). Measures of racism, sexism, heterosexism, and gender binarism for health equity research: From structural injustice to embodied harm – An ecosocial analysis. *Annual Review of Public Health*, *41*(1), 37–62.
Pager, D., & Shepherd, H. (2008). The sociology of discrimination: Racial discrimination in employment, housing, credit, and consumer markets. *Annual Review of Sociology*, *34*(2008), 181–209.
Ralph, L. (2019). The logic of the slave patrol: The fantasy of black predatory violence and the use of force by the police. *Palgrave Communications*, *5*, 130. https://doi.org/10.1057/s41599-019-0333-7.
Roberts, D. R., & Rollins, O. (2020). Why sociology matters to race and biosocial science. *Annual Review of Sociology*, *46*(1), 195–214.
Rosa, J., & Díaz, V. (2020), Raciontologies: Rethinking anthropological accounts of institutional racism and enactments of white supremacy in the United States. *American Anthropologist*, *122*, 120–132. https://doi.org/10.1111/aman.13353.
Schlosser, M. D. (2013). Racial attitudes of police recruits in the United States Midwest Police Academy: A quantitative examination. *International Journal of Criminal Justice Sciences*, *8*(2), 215.
Small, M. L., & Pager, D. (2020). Sociological perspectives on racial discrimination. *Journal of Economic Perspectives*, *34*(2), 49–67.
Swencionis, J. K., & Goff, P. A. (2017). The psychological science of racial bias and policing. *Psychology, Public Policy, and Law*, *23*(4), 398.
Waldman, K. (2014). Demons and Supervillains: The Language of Darren Wilson's Grand Jury Testimony. https://slate.com/human-interest/2014/11/the-language-of-darren-wilson-s-testimony-close-reading-the-demons-and-supervillains.html.

Fighting over who dictates the nature of prejudice

Gordon Hodson

Department of Psychology, Brock University, St. Catharines, Ontario, L2S 3A1, Canada.
ghodson@brocku.ca; www.hodsonlab.com; https://brocku.ca/social-sciences/psychology/people/gordon-hodson/

doi:10.1017/S0140525X21000625, e77

Abstract

A growing trend, reflected in the target article, effectively shifts control of prejudice operationalization to align with right-leaning priorities (and away from disadvantaged groups' voices and social justice). The article would only be compelling if experiments misaligned with real-world findings, if experimenters ignored nuances and moderators, and if the call to consider the social context included the macro-level societal context.

In his provocative article, Cesario challenges the role of experiments in understanding group discrepancies. He is on solid footing when discussing the artificiality of experiments and the field's overconfidence in its gold-standard status as a research tool. His observations are particularly poignant concerning "can-versus-does" interpretations – simply showing that X *can* predict Y does not mean that it *does so* in the real world. These broader methodological concerns, about experiments generally, merit consideration.

Yet his main thesis seems seriously out of touch with the socio-cultural realities and challenges of the twenty-first century, part of a growing trend. As noted by Hodson (2021), the contemporary prejudice field currently risks straying off course given three simultaneous trends: (1) vocalized concerns about concept creep and a desire to *narrow* the operationalization of what constitutes prejudicial attitudes and behavior (e.g., Haslam, 2016); (2) psychological concept *expansion* to include right-wing conceptualizations of "morality" (i.e., ingroup loyalty) (e.g., Haidt & Graham, 2007), plus the dismissal of racial/gender microaggressions as merely subjective and determined by the victim; and (3) declarations of prejudice equivalency (Brandt & Crawford, 2020), such as anti-Black prejudice being equated with anti-banker prejudice, that fail to recognize inherent power and status differentials between groups. These trends risk rendering psychology irrelevant to understanding and repairing societal problems. Cesario's paper represents a fourth column of concern. His ideas would further prioritize the dominant White majority and negate voices from disadvantaged social groups. His advice, if heeded, would delegitimize decades of careful and methodological research on prejudice and discrimination against disadvantaged groups. To what purpose?

At play is control over the narrative concerning the very nature of prejudice – what prejudice *is*. Tellingly, Cesario calls for the field to listen more to the police to understand police shootings, with no mention of listening more to victims (or examining societal factors). At the same time, he paints researchers, mostly (left-leaning) professors, as tricksters who wield omnipotent powers to create artificial worlds that enable them to shape the nature of prejudice. Rather than studying the wider culture wars, this discourse risks playing into them, representing a strong pushback that would prioritize the police academy over the scholarly academy regarding epistemic legitimacy. Worryingly, he objects to the very *idea* of experiments as tools to investigate intergroup inequalities. Here, Cesario misunderstands social psychologists' efforts, who, in unpacking the complexities of prejudice, seek to isolate causes and to discover *whether* "X" can fuel prejudice (not to dictate that X is *the* cause of existing inequalities).

Cesario's case would be more compelling if field experiments and non-experimental work (e.g., archival) contradicted experimental findings. In classic laboratory hiring experiments, qualifications are carefully controlled and made equivalent while group-identifying information (e.g., race and gender) is varied systematically. Using this method, bias against a target can be confidently isolated as group-based or prejudicial. Notably, field experiments show that Whites (vs. Blacks) receive 36% more interview callbacks for interviews (Quillian, Pager, Hexel, & Midtøen, 2017) and 145% more job offers (Quillian, Lee, & Oliver, 2020), consistent with laboratory findings. Large-scale analyses of recruitment platforms, using artificial intelligence to analyze virtually all applicant qualities rather than a single dimension, reveal employment recruiters being 4–19% less likely to contact minority/immigrant candidates relative to Whites (Hangartner, Kopp, & Siegentaler, 2021). In terms of policing, large-scale archival analyses of patrol assignments show White (vs. non-White) officers more likely to stop, arrest, and use force against Black (vs. White) citizens, amplified in non-White neighborhoods (Ba, Knox, Mummolo, & Rivera, 2021). And nationally representative datasets reveal that White police officers, relative to the general population,

https://doi.org/10.1017/S0140525X21000017 Published online by Cambridge University Press

view Black people as violent, express greater racial resentment, and believe that anti-Black discrimination is an historical not contemporary problem (LeCount, 2017). Data from the real-world are thus congruent with those from the experimental paradigm that Cesario criticizes. As such, the bar for disqualifying experiments should be reasonably high, and calls to abolish this methodology should be greeted with healthy skepticism.

His case would also be more compelling if experimentalists failed to consider and contemplate boundary conditions, such as participant type (student vs. police), training/experience effects, cognitive load, and so on. Researchers not only study these nuances but also express clear caution and thoughtfulness. In their review, Payne and Correll (2020) conclude that "while an officer's performance on a laboratory task may provide valuable information, it cannot tell us whether race actually biases decisions about the use of force when police officers encounter suspects in the real world" (p. 36). Cesario's case that experiments create realities incongruent with the real world, and that central researchers extrapolate wildly from laboratory to the real world, are straw-man arguments. Similarly, his calls to consider the bigger context in police shootings would be compelling if he included the macro-level context, including its political and social structures, rather than his limited call to consider the specific micro-level situation (e.g., a specific shooter incident and its lead-up). He wants more information, but not too much.

Cesario's argument fits with a wider trend in academia to control the what-is-prejudice narrative and who gets to decide. As evidenced in the #BlackLivesMatter and #MeToo social movements, disadvantaged and marginalized groups are pleading for more voice at the table, not less. Psychologists express related concerns about the "extreme" and "overwhelming" Whiteness of psychology (see Dupree & Kraus, in press; Roberts, Bereket-Shavit, Dollins, Goldie, & Mortenson, 2020). In a culturally insensitive move, Cesario asks our discipline to direct more causal blame toward shooting victims and troubled children in classrooms, given their supposedly violent and undisciplined natures, for inviting their fates at the hands of the powerful.

As academics, we should be mindful that our ideas and work can be both used and misused. Defence attorneys for George Floyd's killing or the January 6th, 2021 Capitol Hill insurrection will appreciate the intellectual scaffolding these new academic trends offer to the Alt-Right, white supremacists, and those seeking to undo social change and justice. Our discipline lies at a critical crossroads; we can encourage epistemic inclusivity and incorporate more non-White voices, or we can become irrelevant (or detrimental) to the discipline of social studies.

Conflict of interest. No conflicts of interest.

References

Ba, B. A., Knox, D., Mummolo, J., & Rivera, R. (2021). The role of officer race and gender in police-civilian interactions in Chicago. *Science*, *371*(6530), 696–702. doi:10.1126/science.abd8694

Brandt, M. J., & Crawford, J. T. (2020). Worldview conflict and prejudice. *Advances in Experimental Social Psychology*, *61*, 1–66. https://doi.org/10.1016/bs.aesp.2019.09.002.

Dupree, C. H., & Kraus, M. W. (in press). Psychological science is not race neutral. *Perspectives in Psychological Science*. doi:10.1177/1745691620979820

Haidt, J., & Graham, J. (2007). When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize. *Social Justice Research*, *20*(1), 98–116.

Hangartner, D., Kopp, D., & Siegentaler, M. (2021). Monitoring hiring discrimination through online recruitment platforms. *Nature Human Behaviour*, *589*, 572–576. https://doi.org/10.1038/s41586-020-03136-0.

Haslam, N. (2016). Concept creep: Psychology's expanding concepts of harm and pathology. *Psychological Inquiry*, *27*, 1–17. doi:10.1080/1047840X.2016.1082418

Hodson, G. (2021). Pushing back against the microaggression pushback in academic psychology: Reflections on a concept creep paradox. *Perspectives on Psychological Science*, *16*(5) 932–955. DOI: 10.1177/1745691621991863.

LeCount, R. J. (2017). More black than blue? Comparing the racial attitudes of police to citizens. *Sociological Forum*, *32* (S1), 1051–1072. doi:10.1111/socf.12367

Payne, B. K., & Correll, J. (2020). Race, weapons, and the perception of threat. *Advances in Experimental Social Psychology*, *62*, 1–50. https://doi.org/10.1016/bs.aesp.2020.04.001.

Quillian, L., Lee, J. J., & Oliver, M. (2020). Evidence from field experiments in hiring shows substantial additional racial discrimination after the callback. *Social Forces*, *99*(2), 732–759. doi:10.1093/sf/soaa026

Quillian, L., Pager, D., Hexel, O., & Midtøen, A. H. (2017). Meta-analysis of field experiments shows no change in racial discrimination in hiring over time. *Proceedings of the National Academy of Sciences*, *114*(41), 10870–10875. www.pnas.org/cgi/doi/10.1073/pnas.1706255114.

Roberts, S. O., Bereket-Shavit, C., Dollins, F. A., Goldie, P. D., & Mortenson, E. (2020). Racial inequality in psychological research: Trends of the past and recommendations for the future. *Perspectives on Psychological Science*, *15*, 1295–1309. https://doi.org/10.1177/17456916209277.

Missing context from experimental studies amplifies, rather than negates, racial bias in the real world

Leland Jasperse[a], Benjamin S. Stillerman[b] and David M. Amodio[b,c]

[a]Department of English Language and Literature, University of Chicago, Chicago, IL 60637, USA; [b]Department of Psychology, New York University, New York, NY 10003, USA and [c]Department of Psychology, University of Amsterdam, 1001 NK Amsterdam, Netherlands.
ljasperse@uchicago.edu; ben.stillerman@gmail.com; david.amodio@nyu.edu, amodiolab.org

doi:10.1017/S0140525X21000868, e78

Abstract

We agree with Cesario's premise but reject his conclusion: Although experimental studies of racial stereotyping, weapons perception, and shoot decisions typically exclude real-world contextual factors and thus have limited relevance to race disparities (e.g., in policing), these excluded factors comprise systemic, institutional, and individual-level biases that are more likely to amplify racial disparities than negate them.

Cesario claims that experimental findings of racial bias are so disconnected from real-world situations that they "cannot and do not provide information about the nature of group disparities" (sect. 1, para. 2). Indeed, because such experiments are designed to isolate specific cognitive processes, they exclude myriad real-world factors that may otherwise influence intergroup behavior. However, we disagree with Cesario's conclusion that such factors overwhelm effects of social categories like race. In reality, the opposite is true: Real-world situations contain many layers of prejudice and discrimination, typically excluded from lab experiments, and these dramatically compound race effects.

Cesario argues that racial bias is only revealed in experiments when factors such as circumstantial information, group differences, and situational contingencies are omitted. Yet he all but ignores the many powerful layers of systemic, institutional, and

individual racism that pervade real-life interracial interactions. In fact, in U.S. policing, many of the situational factors omitted from lab studies are themselves shaped by race, such as racially motivated profiling and surveillance (Browne, 2015), stop-and-frisk policies (e.g., Cooper, 2018; Gelman, Fagan, & Kiss, 2007; Goel, Rao, & Shroff, 2016), and the use of discriminatory data-driven precision policing (Southerland, 2020). Although Cesario claims these real-world factors "overwhelm [the] strength of categorical bias" (Table 1 of target article), historical and sociological data suggest they actually *exacerbate* group disparities observed in experimental tasks.

To illustrate the supposedly race-neutralizing effect of real-world information, Cesario highlights a study by Correll, Wittenbrink, Park, Judd, and Goyle (2011) but misrepresents the finding. In this modified shooter task, targets are presented in either neutral or "dangerous, urban backgrounds" Cesario writes that the urban background – an instance of "missing information" reintroduced to a task – "completely eliminated racial bias in the decision to shoot" (sect. 4.1.1., para. 4). However, "dangerous, urban" settings are themselves racially coded from decades of segregationist housing policy, racist political rhetoric and media representations, and targeted over-policing (Gordon, 2020; Hurwitz & Peffley, 2005; Rhodes & Brown, 2019). Indeed, the data show that urban backgrounds actually increased the tendency to shoot White targets to the level of Black targets – an unsurprising effect given that these backgrounds themselves contain race-stereotypic cues.

As a real-world illustration, consider the NYPD's killing of Amadou Diallo, a case that galvanized research on implicit bias in shoot decisions: Four white NYPD officers patrolling the Bronx neighborhood of Soundview stopped Diallo, a young Black man "acting suspiciously" who allegedly matched the description of wanted criminal. When Diallo reached into his pocket for his wallet, the lead officer, per his testimony, misidentified it as a gun, triggering the group to shoot and kill Diallo. What other factors were at play that could have overwhelmed the subtle effect of automatic race associations? Notably, the officers were targeting a neighborhood that became majority-Black and over-policed following white flight, economic disinvestment, and redlining (Nonko, 2016; Stoudt, Fine, & Fox, 2011). Moreover, the officers were part of the infamous NYPD Street Crimes Unit, which expressly targeted *dangerous, urban* communities of color to turn up guns and drugs to meet quotas (Harring, 2000). Attributing Diallo's death to a quick decision made in ambiguous circumstances does leave out critical context from this scene, but this context amplifies disparities rather than ameliorates them (Amodio, 2015).

Although studies of implicit bias are often inspired by real-world incidents, they are rarely (if ever) designed to explain them. Instead, they aim to isolate and illuminate basic mechanisms of race processing in the mind; asking, for example, *Can race influence automatic thought and quick decisions?* Such experiments are rarely presented as complete accounts of real-world disparities and expressions of prejudice. Curiously, the article Cesario singles out as "a prototypical example" (sect. 2, para. 2) of this practice, by Moss-Racusin, Dovidio, Brescoll, Graham, and Handelsman (2012), is a field study on gender bias in job applicant evaluations that uses none of the methods he critiques. Moreover, social psychologists have long considered the roles of additional information, forces, and contingencies as moderators of category-based stereotyping (e.g., Amodio & Swencionis, 2018; Darley & Gross, 1983; Dovidio & Gaertner, 2000; Fiske & Neuberg, 1990). The deficiencies Cesario attributes to social psychology appear to concern its translation more than the science itself.

We see a different concern with reductionist experimental studies which, we believe, is much more pressing (Jasperse & Stillerman, 2021): By presenting racial bias as a subtle, unintentional spandrel of the mind, these studies problematically reduce the broad, structural nature of racism to a transient impulse. Consequently, they misdirect efforts toward ineffective training programs (Worden et al., 2020) and give cover to the more pernicious effects of systemic, institutional, and blatant racism. Hence, in addition to underestimating the magnitude of bias, such studies draw attention away from its deeper causes.

Finally, we feel compelled to comment on the selective scholarship and rhetoric in this target article. Cesario elides evidence that racial bias is a pervasive dimension of policing and criminal justice – one that inflects (and exceeds) moment-to-moment individual cognition. He then suggests that observed real-world disparities are due mainly to behavioral differences between groups. For example, he argues that racial disparities in policing may be more a product of different racial groups' criminal tendencies than bias on the part of police officers. Although he hastens to "make no claims about the origin of these group differences" (sect. 1, para. 6), a casual reader could be forgiven for thinking that Cesario believes elevated criminality "might very well be" (sect. 1, para. 6) a trait feature of racial minorities. This rhetorical pattern – to deny the severity of racial bias and then suggestively attribute disparities to individual merits of group members – follows a familiar refrain known to social psychologists as *modern racism*. Regardless of the authors views and intentions, it is concerning to see this device in mainstream scientific discourse.

In summary, we accept Cesario's premise but reject his conclusion; the many real-world factors often missing from sociocognitive experiments of racial bias are themselves the product of systemic, institutional, and individual racism. To the extent real-world factors overwhelm experimentally observed patterns of bias, the effect of racism is likely much stronger.

Financial support. This research received no specific grant from any funding agency, commercial, or not-for-profit sectors.

Conflict of interest. None.

References

Amodio, D. M. (2015). Prejudiced? Me? *New Scientist 227*:26–27.

Amodio, D. M., & Swencionis, J. K. (2018). Proactive control of implicit bias: A theoretical model and implications for behavior change. *Journal of Personality and Social Psychology 115*(2):255–275.

Browne, S. (2015). *Dark matters*. Duke University Press.

Cooper, F. R. (2018). A genealogy of programmatic stop and frisk: The discourse-to-practice circuit. *University of Miami Law Review 73*(1), 1–78.

Correll, J., Wittenbrink, B., Park, B., Judd, C. M., & Goyle, A. (2011). Dangerous enough: Moderating racial bias with contextual threat cues. *Journal of Experimental Social Psychology 47*:184–189.

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

Dovidio, J. F., & Gaertner, S. L. (2000). Aversive racism and selection decisions: 1989 and 1999. *Psychological Science 11*(4):315–319.

Fiske, S. T., & Neuberg, S. L. (1990). A continuum of impression formation, from category-based to individuating processes: Influences of information and motivation on attention and interpretation. *Advances in Experimental Social Psychology 23*:1–74.

Gelman, A., Fagan, J., & Kiss, A. (2007). An analysis of the New York city police department's "stop-and-frisk" policy in the context of claims of racial bias. *Journal of the American Statistical Association 102*(479), 813–823.

Goel, S., Rao, J. M., & Shroff, R. (2016). Precinct or prejudice? Understanding racial disparities in New York city's stop-and-frisk policy. *The Annals of Applied Statistics 10*(1):365–394.

https://doi.org/10.1017/S0140525X21000017 Published online by Cambridge University Press

Gordon, D. (2020). The police as place-consolidators: The organizational amplification of urban inequality. *Law & Social Inquiry 45*(1):1–27.
Harring, S. L. (2000). The Diallo verdict: Another "tragic accident" in New York's war on street crime? *Social Justice 27*(1), 9–18.
Hurwitz, J., & Peffley, M. (2005). Playing the race card in the post-Willie Horton era: The impact of racialized code words on support for punitive crime policy. *Public Opinion Quarterly 69*(1):99–112.
Jasperse, L., & Stillerman, B. (2021). Beyond bias: The case for an abolitionist psychology. Los Angeles Review of Books.
Moss-Racusin, C. A., Dovidio, J. F., Brescoll, V. L., Graham, M. J., & Handelsman, J. (2012). Science faculty's subtle gender biases favor male students. *Proceedings of the National Academy of Sciences 109*(41):16474–16479.
Nonko, E. (2016). Redlining: How one racist, depression-era policy still shapes New York real estate. Brick Underground.
Rhodes, J., & Brown, L. (2019). The rise and fall of the "inner city": Race, space and urban policy in postwar England. *Journal of Ethnic & Migration Studies 45*(17):3243–3259.
Southerland, V. (2020). The intersection of race and algorithmic tools in the criminal legal system. *Maryland Law Review 80*(3), 487–566.
Stoudt, B. G., Fine, M., & Fox, M. (2011). Growing up policed in the age of aggressive policing policies. *New York Law School Law Review 56*:1331–1370.
Worden, R. E., McLean, S. J., Engel, R. S., Cochran, H., Corsaro, N., Reynolds, D., … Isaza, G. T. (2020). The impacts of implicit bias awareness training in the NYPD.

Experiments make a good breakfast, but a poor supper

Jolanda Jetten, Hema Preya Selvanathan, Charlie R. Crimston, Sarah V. Bentley and S. Alexander Haslam

School of Psychology, The University of Queensland, St. Lucia, 4072 QLD, Australia. j.jetten@psy.uq.edu.au; https://psychology.uq.edu.au/profile/2317/jolanda-jetten; h.selvanathan@uq.edu.au; https://psychology.uq.edu.au/profile/7410/hema-preya-selvanathan; c.crimston@uq.edu.au; https://psychology.uq.edu.au/profile/2698/charlie-crimston; s.bentley@uq.edu.au; https://psychology.uq.edu.au/profile/2536/sarah-bentley; a.haslam@uq.edu.au; https://psychology.uq.edu.au/profile/3181/alex-haslam

doi:10.1017/S0140525X21000662, e79

Abstract

Cesario's analysis has three key flaws. First, the focus on whether an effect is "real" (an "effects flaw") overlooks the importance of theory testing. Second, obsession with effects (a "fetishization flaw") sidelines theoretically informed questions about when and why an effect may arise. Third, failure to take stock of cultural and historical context (a "decontextualization flaw") strips findings of meaning.

Cesario provides a number of good reasons why we should be cautious about relying solely on experimental findings to understand the social world around us. While we welcome the focus on experimental validity (after years of focusing more or less exclusively on problems associated with replication and reliability), unfortunately, his own analysis falls foul of some of the problems that it seeks to rectify. There are three specific flaws in his reasoning, and all three are commonly observed in researchers' understanding of what experiments are meant to do and how they should be used.

First, Cesario's analysis misunderstands the purpose of experiments. Their function is not to try as hard as possible to mimic aspects of the world outside the laboratory so that researchers can establish whether a given effect is observable in the world and hence "real" (e.g., whether or not police officers are racially biased). To imagine that they are is to fall prey to an "effects flaw" in which experimental outcomes are privileged over the processes that produce them.

Instead, then, experiments and the evidence they produce are better suited to the task of testing *theories* of human psychology and behaviour. They do this principally by helping us to understand under *what conditions* a given effect is observed, and *what mechanisms* underlie that effect. Indeed, by focusing on effects rather than processes, Cesario's analysis fails to capitalise on the key value of experiments – namely their capacity to support theory development (Haslam & McGarty, 2001; Swann & Jetten, 2017).

This "effects flaw" is not just present in Cesario's analysis, but is a pervasive problem in the social psychological literature. It is perhaps most apparent in reports of the classic studies in social psychology (e.g., Milgram's obedience studies and Zimbardo's Stanford Prison Experiment; see Smith & Haslam, 2017). For instance, because of the "effect flaw" the contribution of Milgram's obedience studies is routinely misunderstood. For the real theoretical value of the work can be seen to lie less in the 65% obedience rate that was observed in the so-called "baseline condition" (the classic effect reported in most textbooks) than in the many variants that Milgram conducted to explore the conditions under which obedience is either far greater or far weaker (see Jetten & Mols, 2014; Reicher, Haslam, & Smith, 2012). To be sure, experimental effects can capture our attention and make the case for much-needed theory development, but without a theoretical focus and grounding, their contribution is unproductively circumscribed.

Second, while we agree that, on its own, experimental evidence is of limited use, we argue that what is needed is a proper analysis of how experimental evidence should be complemented with other forms of evidence. Here, we would argue that experimental evidence should never be considered in isolation, but always in conjunction with data sourced using complementary methods (e.g., field surveys, longitudinal research, and qualitative work). What is more, theory-derived hypotheses need to be examined in a range of different contexts. Unfortunately, although, experimental evidence is too often seen as the "gold (and only) standard" for our field, with evidence gleaned via other means relegated to the margins.

This prioritization of experimental effects contributes to a "fetishization flaw" associated with what Reicher (2000) refers to as methodolatry. As a result of this there is little incentive for researchers to move out of the lab, and once an "effect" is established within a controlled laboratory setting, it hardly ever comes out of it. The experimental paradigm, therefore, becomes equated with the phenomena itself. This exacerbates the consequences of the first flaw by cultivating an obsession with (the replication of) experimental effects and attendant neglect of broader questions of process. In short, questions of "when" and "why" are crowded out by questions of "whether" and "how much" in ways that stymie and suppress theory development and the deep understanding that accompanies it. As the replication crisis of recent years attests, this narrowing of the field has not served social psychology well.

Third, alongside these issues, a "decontextualization flaw" means that researchers typically use experiments for hypothetico-deductive purposes in a quest to discover "objective truth." This epistemology generally assumes value neutrality and context independence and tends to catalogue psychological effects with scant regards to the broader historical and societal contexts in which they arise (Adams, Estrada-Villalta, Sullivan, & Markus, 2019).

In crucial ways, this has led to the disappearance of the "social" in social psychology (see Greenwood, 2003). For it is important to

remember that the underlying causes and nature of systemic issues such as discrimination and inequality cannot be reduced to (or sufficiently captured within) experiments alone. Rather, these realities – and the questions they raise – need to be explored within the worlds that give rise to them (Oishi & Graham, 2010; Trawalter, Bart-Plange, & Hoffman, 2020). Here, qualitative methods are often particularly valuable by virtue of their inductive, reflexive, and phenomenological potential. Critically too, these alternative (and complementary) methodologies are better able to capture the meaning of data in situ and prioritize community participation in the co-creation of knowledge – something which is all too often missing in experimental research (Burman, 1997).

In sum, as with a good breakfast, experiments are an excellent point of departure. But on their own, they can never be enough to satisfy our scientific appetites. For their scientific potential to be fulfilled, their contributions need to be consolidated with meaningful theory development and complementary methodologies. Lacking this, not only will our diet be unbalanced, but it will also be profoundly unsatisfying – and potentially harmful.

Financial support. This research was supported by an Australian Research Council Laureate Fellowship (FL180100094) awarded to Jolanda Jetten.

Conflict of interest. The authors have no conflicts of interests to report in relation to this commentary.

References

Adams, G., Estrada-Villalta, S., Sullivan, D., & Markus, H.R. (2019). The psychology of neoliberalism and the neoliberalism of psychology. *Journal of Social Issues*, *75*, 189–216.

Burman, E. (1997). Minding the gap: Positivism, psychology, and the politics of qualitative methods. *Journal of Social Issues*, *53*(4), 785–801.

Greenwood, J. D. (2003). *The disappearance of the social in American social psychology*. Cambridge University Press.

Haslam, S. A., & McGarty, C. (2001). A hundred years of certitude? Social psychology, the experimental method and the management of scientific uncertainty. *British Journal of Social Psychology*, *40*, 1–21. doi: 10.1348/014466601164669

Jetten, J., & Mols, F. (2014). 50–50 Hindsight: Appreciating anew the contributions of Milgram's obedience experiments. *Journal of Social Issues*, *70*, 587–602.

Oishi, S., & Graham, J. (2010). Social ecology: Lost and found in psychological science. *Perspectives on Psychological Science*, *5*(4), 356–377.

Reicher, S. D. (2000). Against methodolatry. *British Journal of Clinical Psychology*, *39*(1), 1–6.

Reicher, S. D., Haslam, S. A., & Smith, J. R. (2012). Working towards the experimenter: Reconceptualizing obedience within the Milgram paradigm as identification-based followership. *Perspectives on Psychological Science*, *7*, 315–324. doi: 10.1177/1745691612448482

Smith, J. R., & Haslam, S. A. (Eds.) (2017). *Social psychology: Revisiting the classic studies* (2nd ed.). Sage.

Swann, W. B. Jr., & Jetten, J. (2017). Restoring agency to the human actor. *Perspectives on Psychological Science*, *12*, 382–399.

Trawalter, S., Bart-Plange, D. J., & Hoffman, K. M. (2020). A socioecological psychology of racism: Making structures and history more visible. *Current Opinion in Psychology*, *32*, 47–51.

What can the implicit social cognition literature teach us about implicit social cognition?

Benedek Kurdi and Yarrow Dunham

Department of Psychology, Yale University New Haven, CT 06511.
benedek.kurdi@yale.edu; http://www.benedekkurdi.com
yarrow.dunham@yale.edu; http://www.socialcogdev.com

doi:10.1017/S0140525X21000595, e80

Abstract

We highlight several sets of findings from the past decade elucidating the relationship between implicit social cognition and real-world inequality: Studies focusing on practical ramifications of implicit social cognition in applied contexts, the relationship between implicit social cognition and consequential real-world outcomes at the level of individuals and geographic units, and convergence between individual-level and corpus-based measures of implicit bias.

The target article calls for "systematically dismantling" the "fundamentally flawed" practice of using implicit social cognition research to inform our understanding of real-world inequality. Sweeping conclusions and comprehensive recommendations of this kind, published in a leading journal of our discipline, should be supported by powerful arguments reflecting the latest state of the literature. Instead, the target article mischaracterizes the methods, goals, and state of implicit social cognition research while referencing a mere eight empirical papers, the most recent of which was published over a decade ago.

According to the target article, the "standard research cycle" begins with the observation that groups differ on some real-world outcome and has the goal of explaining, and eventually eliminating, such differences. This statement is misleadingly narrow. Not all memory research seeks to cure dementia; not all phonological awareness research tries to eradicate dyslexia; and not all auditory perception research contributes to the development of hearing aids. Similarly, much implicit social cognition research explores basic aspects of thought and behavior, including learning and representation (Kurdi & Dunham, 2020), social cognitive development (Dunham, Baron, & Banaji, 2008), and cultural change (Charlesworth & Banaji, 2019), without making any claim of immediate applicability to real-world problems. Thus, whether implicit social cognition research can explain real-world inequality should not be treated as its sole measure of success.

Of course, some of this literature does speak to real-world outcomes and behaviors. But here too the target article misses the mark. Specifically, according to the target article, researchers establish some experimental effect of social category knowledge in a small sample of naïve undergraduate participants in the lab and, without any further ado, conclude that the processes observed in the lab directly explain real-world disparities. In fact, as discussed below, much recent implicit social cognition research does not bear much resemblance to this description.

One relevant line of research has documented practical ramifications of basic implicit cognitive processes. For instance, transgender and cisgender children have been shown not to meaningfully differ from each other in implicit gender identity (Olson, Key, & Eaton, 2015), thus providing a counterweight to prior claims of "psychological deviance." In other cases, changes in implicit social cognition have been shown to track meaningful experiences in field settings: For example, exposure to female college professors in science, technology, engineering, and mathematics (STEM) fields can produce long-term effects on implicit gender stereotypes and self-concept (Dasgupta & Asgari, 2004), implying that the social structures in which we are embedded shape the ways in which we envision our future possibilities.

Other research has investigated the relationship between implicit measures and ecologically meaningful measures of intergroup behavior (Kurdi et al., 2019b). For example, implicit math–gender stereotypes predict actual academic achievement among high school students (Steffens, Jelenec, & Noack, 2010); implicit weight stereotypes predict actual callbacks of job applicants among human resources professionals (Agerström & Rooth, 2011); managers' implicit competence stereotypes predict actual job performance of their minority employees (Glover, Pallais, & Pariente, 2017); and doctors' implicit evaluations predict actual rapport, satisfaction, and treatment adherence among Black patients (Hagiwara et al., 2013; Penner et al., 2016, 2010).

Echoing an oft-repeated argument, the target article hastens to underscore that studies of predictive validity produce small correlations between implicit attitudes and intergroup behavior. The finding that the relationship between explicit attitudes and intergroup behavior is almost exactly the same size (Kurdi et al., 2019b) receives no mention. What's more, the mean implicit–behavior correlation sits right around the 25th percentile of all effect sizes in social psychology, with the largest implicit–behavior correlations at the individual level approaching the 70th percentile of that distribution (Lovakov & Agadullina, 2021).

Equally absent is any discussion of studies that investigate the association between implicit cognition and real-world inequality at the level of geographic units, which have produced large effects in multiple domains (Hehman, Calanchini, Flake, & Leitner, 2019; Payne, Vuletich, & Lundberg, 2017). For example, this work has demonstrated that regions with higher levels of implicit race bias are characterized by more frequent police killings of Black Americans (Hehman, Flake, & Calanchini, 2018), as well as more racial disparity in school disciplinary actions (Riddle & Sinclair, 2019) and upward mobility (Chetty, Hendren, Jones, & Porter, 2020).

Finally, remarkable correspondence has also been found between individual-level conceptual associations indexed by implicit measures and cultural-level conceptual associations computationally derived from vast amounts of text produced spontaneously and outside any experimental setting (Caliskan & Lewis, 2020). Evidence for such alignment has been provided across different contexts, including a comprehensive examination of social group attitudes and stereotypes (Caliskan, Bryson, & Narayanan, 2017), the relationship between implicit beliefs and evaluations (Kurdi, Mann, Charlesworth, & Banaji, 2019a), and the development of gender biases over the lifespan (Charlesworth, Yang, Mann, Kurdi, & Banaji, 2021).

Little, if any, of the criticism formulated in the target article seems applicable to methodologically sound implicit social cognition research conducted over the past decade. Far from simply assuming a one-to-one correspondence between findings obtained with small undergraduate samples in artificial lab settings and real-world inequality, an increasingly large group of investigators have made serious efforts to establish connections between implicit measures of social cognition and group-based disparities. Specifically, all of the studies discussed above include at least one (but typically all) of the following elements: samples consisting of experts or members of the general public; real behaviors of consequence observed under ecologically realistic conditions; and the availability of ample individuating information during the decision-making process.

Implicit social cognition research has obviously not been immune to some of the same methodological missteps that have troubled much of psychology and the behavioral sciences over the past few decades. However, as should be clear based on even this brief review, there is considerable reason for optimism. Most importantly, further improvement and innovation won't be fueled by throwing out the baby with the bathwater. Instead, whether the goal is basic science or uncovering the antecedents, mechanisms, and consequences of real-world inequality, we urge renewed focus on theory building, study design, and statistical inference. And accurately characterizing the field that one critiques.

Financial support. This research received no specific grant from any funding agency, commercial, or not-for-profit sectors.

Conflict of interest. Benedek Kurdi is a member of the Scientific Advisory Board of Project Implicit, a 501(c)(3) non-profit organization and international collaborative of researchers who are interested in implicit social cognition.

References

Agerström, J., & Rooth, D.-O. (2011). The role of automatic obesity stereotypes in real hiring discrimination. *Journal of Applied Psychology*, *96*(4), 790–805. http://doi.org/10.1037/a0021594.

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183–186. http://doi.org/10.1126/science.aal4230.

Caliskan, A., & Lewis, M. (2020). Social biases in word embeddings and their relation to human cognition. *PsyArXiv*. http://doi.org/10.31234/osf.io/d84kg.

Charlesworth, T. E. S., & Banaji, M. R. (2019). Patterns of implicit and explicit attitudes: I. Long-term change and stability from 2007 to 2016. *Psychological Science*, *30*(2), 174–192. http://doi.org/10.1177/0956797618813087.

Charlesworth, T. E. S., Yang, V., Mann, T. C., Kurdi, B., & Banaji, M. R. (2021). Gender stereotypes in natural language: Word embeddings show robust consistency across child and adult language corpora of more than 65 million words. *Psychological Science*, *Psychological Science*, *32*(2), 218–240. http://doi.org/10.1177/0956797620963619.

Chetty, R., Hendren, N., Jones, M. R., & Porter, S. R. (2020). Race and economic opportunity in the United States: An intergenerational perspective. *The Quarterly Journal of Economics*, *135*(2), 711–783. http://doi.org/10.1093/qje/qjz042.

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*(5), 642–658. http://doi.org/10.1016/j.jesp.2004.02.003.

Dunham, Y., Baron, A. S., & Banaji, M. R. (2008). The development of implicit intergroup cognition. *Trends in Cognitive Sciences*, *12*(7), 248–253. http://doi.org/10.1016/j.tics.2008.04.006.

Glover, D., Pallais, A., & Pariente, W. (2017). Discrimination as a self-fulfilling prophecy: Evidence from French grocery stores. *The Quarterly Journal of Economics*, *132*(3), 1219–1260. http://doi.org/10.1093/qje/qjx006.

Hagiwara, N., Penner, L. A., Gonzalez, R., Eggly, S., Dovidio, J. F., Gaertner, S. L. (2013). Racial attitudes, physician-patient talk time ratio, and adherence in racially discordant medical interactions. *Social Science & Medicine*, *87*(C), 123–131. http://doi.org/10.1016/j.socscimed.2013.03.016.

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

Hehman, E., Flake, J. K., & Calanchini, J. (2018). Disproportionate use of lethal force in policing is associated with regional racial biases of residents. *Social Psychological and Personality Science*, *9*(4), 393–401. http://doi.org/10.1177/1948550617711229.

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

Kurdi, B., Mann, T. C., Charlesworth, T. E. S., & Banaji, M. R. (2019a). The relationship between implicit intergroup attitudes and beliefs. *Proceedings of the National Academy of Sciences*, *21*, 201820240–10. http://doi.org/10.1073/pnas.1820240116.

Kurdi, B., Seitchik, A. E., Axt, J. R., Carroll, T. J., Karapetyan, A., Kaushik, N. (2019b). Relationship between the implicit association test and intergroup behavior: A meta-analysis. *American Psychologist*, *74*(5), 569–586. http://doi.org/10.1037/amp0000364.

Lovakov, A., & Agadullina, E. R. (2021). Empirically derived guidelines for effect size interpretation in social psychology. *European Journal of Social Psychology*, *51*(3), 485–504. http://doi.org/10.1002/ejsp.2752.

Olson, K. R., Key, A. C., & Eaton, N. R. (2015). Gender cognition in transgender children. *Psychological Science*, *26*(4), 467–474. http://doi.org/10.1177/0956797614568156.

https://doi.org/10.1017/S0140525X21000017 Published online by Cambridge University 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*(4), 233–248. http://doi.org/10.1080/1047840X.2017.1335568.

Penner, L. A., Dovidio, J. F., Gonzalez, R., Albrecht, T. L., Chapman, R., Foster, T. … Eggly, S. (2016). The effects of oncologist implicit racial bias in racially discordant oncology interactions. *Journal of Clinical Oncology*, *34*(24), 2874–2880. http://doi.org/10.1200/JCO.2015.66.3658.

Penner, L. A., Dovidio, J. F., West, T. V., Gaertner, S. L., Albrecht, T. L., Dailey, R. K., & Markova, T. (2010). Aversive racism and medical interactions with black patients: A field study. *Journal of Experimental Social Psychology*, *46*(2), 436–440. http://doi.org/10.1016/j.jesp.2009.11.004.

Riddle, T., & Sinclair, S. (2019). Racial disparities in school-based disciplinary actions are associated with county-level rates of racial bias. *Proceedings of the National Academy of Sciences*, *116*(17), 8255–8260. http://doi.org/10.1073/pnas.1808307116.

Steffens, M. C., Jelenec, P., & Noack, P. (2010). On the leaky math pipeline: Comparing implicit math–gender stereotypes and math withdrawal in female and male children and adolescents. *Journal of Educational Psychology*, *102*(4), 947–963. http://doi.org/10.1037/a0019920.

The unbearable limitations of solo science: Team science as a path for more rigorous and relevant research

Alison Ledgerwood[a], Cynthia Pickett[b], Danielle Navarro[c], Jessica D. Remedios[d] and Neil A. Lewis Jr.[e]

[a]Department of Psychology, University of California, Davis, CA 95616, USA; [b]Office of the Provost, DePaul University, Chicago, IL 60604, USA; [c]Department of Psychology, University of New South Wales, 2052 Sydney, Australia; [d]Department of Psychology, Tufts University, Medford, MA 02155, USA and [e]Department of Communication, Cornell University, Ithaca, NY 14853, USA.
aledgerwood@ucdavis.edu
cindy.pickett@depaul.edu
d.navarro@unsw.edu.au
Jessica.Remedios@tufts.edu
nlewisjr@cornell.edu
http://www.alisonledgerwood.com/
https://csh.depaul.edu/faculty-staff/faculty-a-z/Pages/psychology/cynthia-pickett.aspx
https://djnavarro.net/
https://as.tufts.edu/psychology/social-identity-and-stigma-lab
https://neillewisjr.com/

doi:10.1017/S0140525X21000844, e81

Abstract

Both early social psychologists and the modern, interdisciplinary scientific community have advocated for diverse team science. We echo this call and describe three common pitfalls of solo science illustrated by the target article. We discuss how a collaborative and inclusive approach to science can both help researchers avoid these pitfalls and pave the way for more rigorous and relevant research.

In 1946, Lewin wrote about the importance of conducting "action research" that could improve intergroup relations. Lewin and his contemporaries recognized that to do action research well, psychologists could not work alone. To do so would limit their ability to answer three critical questions regarding the phenomenon under study: "(1) What is the present situation? (2) What are the dangers? (3) And most important of all, what shall we do?" (Lewin, 1946, p. 34). They learned that rigorous and relevant social psychological research requires collaborating not only with scientists in other disciplines to understand the full range of forces acting upon a person in a social system, but also with community partners, governments, and other local stakeholders who have direct access to information and insights about how those forces operate in the specific context at hand (IJzerman et al., 2020). Indeed, a growing consensus across disciplines recognizes the value of a collaborative, multidisciplinary, and inclusive approach to science (Albornoz, Posada, Okune, Hillyer, & Chan, 2017; Disis & Slattery, 2010; Ledgerwood et al., 2021; Murphy et al., 2020).

The importance of a collaborative approach was well-known in the early days of psychology but has been neglected in the modern era (Cialdini, 2009). Neglecting the true *powers of the situation* the cultural, economic, historical, political, and sociological forces that affect the mind (including the minds of psychologists) limits the rigor and relevance of the discipline's research, and hampers psychologists' ability to truly understand the conditions under which our work is or is not relevant for social issues.

In his target article, Cesario discusses challenges he perceives in social psychological experiments on bias, and concludes that we should abandon such experiments. While we agree that many experiments have flaws, our view is that Cesario's own critique suffers from three flaws that render his conclusion premature (Table 1). We further suggest that these flaws could have been avoided by collaborating with multidisciplinary experts or even experts in other areas of psychology.

The first flaw is the *biased search flaw*: When people's expectations lead them to consider an incomplete set of possibilities or to search through available information in a manner shaped by personal expectations (Cameron & Trope, 2004). This flaw is costly because it leads to mistaken conclusions based on an incomplete survey of possible alternatives. For example, the target article correctly notes that effect sizes depend on the paradigm used to study them (Kennedy, Simpson, & Gelman, 2019; McShane & Böckenholt, 2014). However, it discusses only the possibility that effect sizes observed in the lab would diminish in the world, and omits the possibility that they would be magnified. After all, in the real world, effects of discrimination compound over time (Krieger & Sidney, 1996; Mays, Cochran, & Barnes, 2007); small effects can become large when compounded across many decisions (Funder & Ozer, 2019). Similarly, although lab studies typically only manipulate a single dimension of bias, in the world, dimensions of bias can intersect to produce compounded or unique effects (Berdahl & Moore, 2006; Remedios & Sanchez, 2018; Settles & Buchanan, 2014). Moreover, research suggests that biases can be magnified when people have access to rich information (as in the real world) that can be marshaled to elaborate and rationalize initial expectations (Darley & Gross, 1983; Taber & Lodge, 2006).

The second flaw is the *beginner's bubble flaw*: when people know a little about a topic but overestimate how well they understand it (Sanchez & Dunning, 2018). This flaw is costly because it leads scholars to misapply or miss insights developed in other areas. For example, the target article relies heavily on the idea that in the real world, people use information that "may be probabilistically accurate in everyday life" (sect. 5, para. 7) and that using demographic information (e.g., race) to fill in the blanks when full information is unavailable is rational in a Bayesian sense and therefore unbiased. This vague and imprecise assertion muddies waters that

Table 1 (Ledgerwood et al.). Three common flaws in solo science illustrated by the target article

Flaw	Description	How diverse team science can help
The biased search flaw	When scholars' expectations lead them to consider an incomplete set of possibilities or to search through available information in a way that is shaped by what they personally expect to find.	By working in teams that include scholars from diverse vantage points, scholars are more likely to encounter and consider different expectations and possibilities.
The beginner's bubble flaw	When scholars know a little bit about a topic but overestimate how well they understand it.	By working with experts in different areas, scholars can leverage each other's deep expertise in specific areas to complement their own. Collaborating with experts in other areas also provides a useful check on whether we understand an area as well as we think we do.
The old wine in new bottles flaw	When scholars (often unintentionally) approach a well-studied idea without recognizing relevant prior work.	A team of diverse collaborators can pool their expertise to create a more comprehensive and generative set of connections to relevant work across disciplinary boundaries.

have already been clarified at length in adjacent literatures, including in-depth discussions by cognitive modelers on the limits of Bayesian theorizing (Bowers & Davis, 2012; Jones & Love, 2011) and clear distinctions between truth and bias developed in social psychological models of judgment (West & Kenny, 2011). Even advocates of Bayesian cognitive models do not claim a behavior is rational or justifiable simply by virtue of being Bayesian (Griffiths, Chater, Norris, & Pouget, 2012; Tauber, Navarro, Perfors, & Steyvers, 2017). A prior is not the same thing as a base rate, nor is it the same thing as truth (Welsh & Navarro, 2012). Just because a belief can *sometimes* lead to correct decisions does not mean it is accurate or optimal to use that belief for all decisions.

The third flaw is the *old wine in new bottles flaw*: when scholars approach a well-studied idea without recognizing relevant prior work. This flaw is costly because it impedes cumulative and integrative science. For example, discussions of how to connect the world and the lab can and should be grounded in the rich, interdisciplinary work on these questions (Aronson & Carlsmith, 1968; Bauer, Damschroder, Hagedorn, Smith, & Kilbourne, 2015; IJzerman et al., 2020; Lewin, 1946; Premachandra & Lewis, 2021). Similarly, previous discussions of external validity have inspired considerable research that helpfully spans the "troubling…gap" (p. 42) between highly controlled studies of bias and disparate treatment in complex real-world contexts (e.g., Dupas, Modestino, Niederle, & Wolfers, 2021; Sarsons, 2017).

These three flaws illustrate common pitfalls for researchers who attempt to tackle large and complex problems from a single vantage point, but they can be mitigated or avoided by working collaboratively in diverse teams (Ledgerwood et al., 2021; Murphy et al., 2020). The key to successfully connecting the lab with the real world is not to abandon experiments on socially relevant topics, but instead for social psychologists to form collaborative partnerships with organizations that can provide on-the-ground insights that lead us to design *better* experiments (IJzerman et al., 2020).

Acknowledgements. This study was supported by NSF #BCS-1941440 to A.L. and a Faculty Fellowship from the Cornell Center for Social Sciences to N.L. The authors thank Katherine Weltzien, Paul Eastwick, Srilaxmi Pappoppula, Stephanie Goodwin, and Sylvia Liu for their help.

Conflict of interest. None.

References

Albornoz, D., Posada, A., Okune, A., Hillyer, R., & Chan, L. (2017). Co-constructing an open and collaborative manifesto to reclaim the open science narrative. *Expanding Perspectives on Open Science: Communities, Cultures and Diversity in Concepts and Practices*, 293–304.

Aronson, E., & Carlsmith, J. M. (1968). Experimentation in social psychology. In G. Lindzey & E. Aronson (Eds.), *The handbook of social psychology* (2nd ed., Vol. 2, pp. 1–79). Addison-Wesley.

Bauer, M. S., Damschroder, L., Hagedorn, H., Smith, J., & Kilbourne, A. M. (2015). An introduction to implementation science for the non-specialist. *BMC Psychology 3* (1):1–12.

Berdahl, J. L., & Moore, C. (2006). Workplace harassment: Double jeopardy for minority women. *Journal of Applied Psychology 91*(2):426–436.

Bowers, J. S., & Davis, C. J. (2012). Bayesian just-so stories in psychology and neuroscience. *Psychological Bulletin 138*:389–414.

Cameron, J. A., & Trope, Y. (2004). Stereotype-biased search and processing of information about group members. *Social Cognition 22*:650–672.

Cialdini, R. B. (2009). We have to break up. *Perspectives on Psychological Science 4*:5–6.

Darley, J. M., & Gross, P. H. (1983). A hypothesis-confirming bias in labeling effects. *Journal of Personality and Social Psychology 44*:20.

Disis, M. L., & Slattery, J. T. (2010). The road we must take: Multidisciplinary team science. *Science Translational Medicine 2*(22):22cm9–22cm9.

Dupas, P., Modestino, A. S., Niederle, M., & Wolfers, J. (2021). *Gender and the dynamics of economics seminars (No. w28494)*. National Bureau of Economic Research.

Funder, D. C., & Ozer, D. J. (2019). Evaluating effect size in psychological research: Sense and nonsense. *Advances in Methods and Practices in Psychological Science 2*:156–168.

Griffiths, T. L., Chater, N., Norris, D., & Pouget, A. (2012). How the Bayesians got their beliefs (and what those beliefs actually are): Comment on Bowers and Davis (2012). *Psychological Bulletin 138*:415–422.

IJzerman, H., Lewis, N. A., Przybylski, A. K., Weinstein, N., DeBruine, L., Ritchie, S. J., … Anvari, F. (2020). Use caution when applying behavioural science to policy. *Nature Human Behaviour 4*(11):1092–1094.

Jones, M., & Love, B. C. (2011). Bayesian fundamentalism or enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition. *Behavioral and Brain Sciences 34*:169–188.

Kennedy, L., Simpson, D., & Gelman, A. (2019). The experiment is just as important as the likelihood in understanding the prior: A cautionary note on robust cognitive modeling. *Computational Brain & Behavior 2*(3):210–217.

Krieger, N., & Sidney, S. (1996). Racial discrimination and blood pressure: The CARDIA study of young black and white adults. *American Journal of Public Health 86* (10):1370–1378.

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

Lewin, K. (1946). Action research and minority problems. *Journal of Social Issues 2* (4):34–46.

Mays, V. M., Cochran, S. D., & Barnes, N. W. (2007). Race, race-based discrimination, and health outcomes among African Americans. *Annual Review of Psychology 58*:201–225.

McShane, B. B., & Böckenholt, U. (2014). You cannot step into the same river twice: When power analyses are optimistic. *Perspectives on Psychological Science 9*:612–625.

Murphy, M. C., Mejia, A. F., Mejia, J., Yan, X., Cheryan, S., Dasgupta, N., … Pestilli, F. (2020). Open science, communal culture, and women's participation in the movement to improve science. *Proceedings of the National Academy of Sciences 117*(39):24154–24164.

Premachandra, B., & Lewis Jr N. (2021). Do we report the information that is necessary to give psychology away? A scoping review of the psychological intervention literature

2000–2018. *Perspectives on Psychological Science*, *17*(1). https://doi.org/10.1177/1745691620974774.
Remedios, J. D., & Sanchez, D. T. (2018). Intersectional and dynamic social categories in social cognition. *Social Cognition 36*:453–460.
Sanchez, C., & Dunning, D. (2018). Overconfidence among beginners: Is a little learning a dangerous thing? *Journal of Personality and Social Psychology 114*:10–28.
Sarsons, H. (2017). Recognition for group work: Gender differences in academia. *American Economic Review 107*:141–145.
Settles, I. H., & Buchanan, N. T. (2014). Multiple groups, multiple identities, and intersectionality. In V. Benet-Martínez & Y.-Y. Hong (Eds.), *Oxford Library of psychology. The Oxford handbook of multicultural identity* (pp. 160–180). Oxford University Press.
Taber, C. S., & Lodge, M. (2006). Motivated skepticism in the evaluation of political beliefs. *American Journal of Political Science 50*:755–769.
Tauber, S., Navarro, D. J., Perfors, A., & Steyvers, M. (2017). Bayesian Models of cognition revisited: Setting optimality aside and letting data drive psychological theory. *Psychological Review 124*:410–441.
Welsh, M. B., & Navarro, D. J. (2012). Seeing is believing: Priors, trust, and base rate neglect. *Organizational Behavior and Human Decision Processes 119*(1):1–14.
West, T. V., & Kenny, D. A. (2011). The truth and bias model of judgment. *Psychological Review 118*:357–378.

Missing perspective: Marginalized groups in the social psychological study of social disparities

Jes L. Matsick, Flora Oswald and Mary Kruk

Departments of Psychology and Women's, Gender, and Sexuality Studies, The Pennsylvania State University, University Park, PA 16803, USA.
jmatsick@psu.edu; https://jmatsick.wixsite.com/uplab
feo5020@psu.edu;
mxk724@psu.edu

doi:10.1017/S0140525X21000601, e82

Abstract

Drawing on interdisciplinary, feminist insights, we encourage social psychologists to embrace the active participation of marginalized groups in social disparities research. We explain (1) how the absence of marginalized groups' perspectives in research presents a serious challenge to understanding intergroup dynamics and concomitant disparities, and (2) how their inclusion could assuage some of social psychology's "fatal flaws."

Cesario argued that three flaws permeate social psychology and undermine what psychologists know about social disparities, but Cesario has not thoroughly acknowledged a potential solution to draw upon: the perspectives and experiences of marginalized groups (e.g., those disadvantaged by gender, race, and sexual orientation). The relative absence of marginalized groups' perspectives in social psychological research presents a serious challenge to understanding intergroup dynamics and concomitant disparities, and their inclusion may offer an antidote to some of the "fatal flaws." We disagree with Cesario about the extent to which the flaws fabricate disparities, but that is not the central claim we take to task. Instead, we advocate for missing perspectives (i.e., marginalized groups' perspectives), which yield benefits for addressing Cesario's concerns and bolstering social psychologists' understanding of disparities.

Social psychological research often identifies what dominant groups do or don't do and touts those findings as evidence for or against social disparities. Given convenience sampling procedures overrepresent dominant groups (Rad, Martingano, & Ginges, 2018), marginalized groups remain relatively neglected in psychological research despite intergroup relations being bidirectional (Roberts, Bareket-Shavit, Dollins, Goldie, & Mortenson, 2020; Shelton, 2000). For example, only 5% of articles in one premier psychology journal predominately sampled U.S. ethnic minorities (Thalmayer, Toscanelli, & Arnett, 2020), and less than 2% of psychological studies across three decades of research included sexual minorities as participants (Lee & Crawford, 2012). As in Cesario, researchers often position marginalized groups as experimental stimuli upon which to be acted; however, beyond their roles as targets, marginalized groups add value to the study of disparities as informants of intergroup relations (Shelton, 2000). Feminist standpoint theory offers a framework for appreciating the advantages of marginalized groups' perspectives in research. It stresses that knowledge is situated, marginalization privies low-status groups to knowledge that is unavailable to dominant groups, and research about power should prioritize those most marginalized (Crasnow, 2020; Haraway, 1988; Harding, 2004; Rolin, 2009). By centering marginalized groups, social psychologists will improve their science of social disparities and remedy extant limitations.

We first consider Cesario's *missing information flaw* (i.e., experiments remove information that is valuable in real-world scenarios). Research that begins by probing marginalized groups' experiences can identify relevant features of real-world situations to retain for lab-based studies. For example, sexual minorities indicate that their experiences of discrimination rely on gender expression (i.e., the extent to which they "pass" as heterosexual and as conventionally feminine/masculine); however, when social psychologists assess sexual stigma, rarely do they manipulate target gender expression despite sexual minorities reporting that people use information about their gender expression to enact bias (e.g., Anderson, 2020; Hoskin, 2019). Consistent with standpoint theory, marginalized groups may possess superior awareness of inequality and injustice. Although members of dominant groups may not discern which sources of information exacerbate bias, members of marginalized groups may more easily notice the circumstances under which bias occurs.

Second, the inclusion of marginalized groups as participants should address Cesario's concern over *missing forces*: the absence of marginalized groups' behaviors in experiments. Research on intergroup interactions provides exemplary support for marginalized groups' inclusion in research. Such an approach empowers marginalized groups as active agents in the research process beyond being passive targets of dominant groups' actions (Shelton, 2000). It also fosters a bidirectional account of intergroup dynamics, which answers Cesario's call for lab-based studies to account for the role of marginalized groups. We propose that their *real* presence in research may increase bias. For example, although intergroup anxiety emerges in intergroup interactions, stressors differ. Dominant groups worry about appearing likeable and non-prejudiced, whereas marginalized groups worry about stigma (Shelton, 2003). Given that real intergroup interactions evoke stress, anxiety, and misunderstanding (MacInnis & Page-Gould, 2015; Richeson & Shelton, 2007; Schultz, Gaither, Urry, & Maddox, 2015; Vorauer, 2006), intergroup exchanges can produce negative consequences (e.g., heightened ingroup favoritism and avoidance of future contact) –

revealing biased processes not as easily captured by research using hypothetical, imagined outgroup members that induce relatively less anxiety.

Third, social psychologists often overlook marginalized groups' expert, first-hand knowledge of disparities, which could address some of Cesario's *missing contingencies*. By adopting person-centered, intersectional approaches, social psychologists could highlight within-group and intergroup variance in how people interpret bias (e.g., Carter & Murphy, 2015; Eibach & Ehrlinger, 2006). This information would prove useful for addressing Cesario's assertion that biases are not uniformly experienced. Intersectional approaches also necessitate an understanding of multiple, interlocking social identities and social systems for addressing inequalities (Crenshaw, 1989, 1991; hooks, 1984), including contingencies of people's other social positions. As an example, biases in science, technology, engineering, and mathematics (STEM)-based evaluation involve more complexity than a gender effect and thus should be considered multidimensionally (Eaton, Saunders, Jacobson, & West, 2020). Indeed, Black women experience sexism in ways inextricably linked to racism, whereas White women's experiences of sexism dovetail with White privilege (Bowleg, 2008). Taking intersectionality seriously (see McCormick-Huhn, Warner, Settles, & Shields, 2019) would reintroduce some of Cesario's *missing contingencies* into the social psychological study of disparities.

We also encourage psychologists to travel beyond disciplinary boundaries to appreciate the sociopolitical and historical contexts surrounding disparities they aim to understand. Interdisciplinary consultation with non-psychologists (e.g., feminist scholars, critical race theorists, and humanists) provides rich contextualization of psychological questions and findings (Bowleg, 2008; Grzanka, 2018; Held, 2020; Warner, 2008). For example, embracing humanistic ideals of empathy and subjectivity could transform social psychological questions, such as not only asking "Are shooters biased?," but also "Do Black individuals detect bias when encountering police under differing conditions, and how are Black people psychologically affected by the threat that they anticipate?" Interdisciplinary insights would also help social psychologists connect contemporary research questions to the cultural, historical, and political origins that make such inquiries worthwhile (e.g., connections between slave patrols and modern-day policing; Reichel, 1988).

We remain optimistic that we can build upon social psychological approaches to strengthen the field's scientific contributions, but it requires careful, deliberate attention to marginalized groups' experiences. Increasing social psychologists' attention to marginalized groups responds to Cesario's flaws, enriches the study of social disparities, and diversifies sample representation within psychology. Moving beyond disciplinary lines, social psychologists would benefit from engaging feminist standpoint theory and respecting interdisciplinary knowledge.

Financial support. The author(s) received no financial support for the research, authorship, and/or publication of this article.

Conflict of interest. The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

References

Anderson, S. M. (2020). Gender matters: The perceived role of gender expression in discrimination against cisgender and transgender LGBQ individuals. *Psychology of Women Quarterly*, *44*(3), 323–341. https://doi.org/10.1177/0361684320929354.

Bowleg, L. (2008). "When Black + lesbian + woman ≠ Black lesbian woman": The methodological challenges of qualitative and quantitative intersectionality research. *Sex Roles*, *59*(5–6), 312–325. https://doi.org/10.1007/s11199-008-9400-z.

Carter, E. R., & Murphy, M. C. (2015). Group-based differences in perceptions of racism: What counts, to whom, and why? *Social and Personality Psychology Compass*, *9*, 269–280. https://doi.org/10.1111/spc3.12181.

Crasnow, S. (2020). Feminist perspectives on science. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy (winter 2020 ed.)*. Metaphysics Research Lab, Stanford University. Retrieved March 6, 2020, from: https://plato.stanford.edu/archives/win2020/entries/feminist-science/.

Crenshaw, K. (1989). Demarginalizing the intersection of race and sex: A Black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. *The University of Chicago Legal Forum*, *140*, 139–167.

Crenshaw, K. (1991). Mapping the margins: Intersectionality, identity politics, and violence against women of color. *Stanford Law Review*, *43*, 1241–1299.

Eaton, A. A., Saunders, J. F., Jacobson, R. K., & West, K. (2020). How gender and race stereotypes impact the advancement of scholars in STEM: Professors' biased evaluations of physics and biology post-doctoral candidates. *Sex Roles*, *82*, 127–141. https://doi.org/10.1007/s11199-019-01052-w.

Eibach, R. P., & Ehrlinger, J. (2006). "Keep your eyes on the prize": Reference points and racial differences in assessing progress toward equality. *Personality and Social Psychology Bulletin*, *32*(1), 66–77. https://doi.org/10.1177/0146167205279585.

Grzanka, P. R. (2018). Intersectionality and feminist psychology: Power, knowledge, and process. In C. B. Travis, J. W. White, A. Rutherford, W. S. Williams, S. L. Cook, & K. F. Wyche (Eds.), *APA Handbook of the psychology of women* (pp. 585–602). American Psychological Association.

Haraway, D. (1988). Situated knowledges: The science question in feminism and the privilege of partial perspective. *Feminist Studies*, *14*(3), 575–599. https://doi.org/10.2307/3178066.

Harding, S. (2004). Introduction: Standpoint theory as a site of political, philosophic, and scientific debate. In S. Harding (Ed.), *The feminist standpoint theory reader: Intellectual and political controversies* (pp. 1–16). Routledge.

Held, B. S. (2020). Taking the humanities seriously. *Review of General Psychology*. *25*(2), 119–133. https://doi.org/10.1177/1089268020975024.

hooks, B. (1984). *Feminist theory: From margin to center*. Pluto Press.

Hoskin, R. A. (2019). Femmephobia: The role of anti-femininity and gender policing in LGBTQ + people's experiences of discrimination. *Sex Roles*, *81*, 686–703. https://doi.org/10.1007/s11199-019-01021-3.

Lee, I.-C., & Crawford, M. (2012). Lesbians in empirical psychological research: A new perspective for the twenty-first century? *Journal of Lesbian Studies*, *16*(1), 4–16. https://doi.org/10.1080/10894160.2011.557637.

MacInnis, C. C., & Page-Gould, E. (2015). How can intergroup interaction be bad if intergroup contact is good? Exploring and reconciling an apparent paradox in the science of intergroup relations. *Perspectives on Psychological Science*, *10*(3), 307–327. https://doi.org/10.1177/1745691614568482.

McCormick-Huhn, K., Warner, L. R., Settles, I. H., & Shields, S. A. (2019). What if psychology took intersectionality seriously? Changing how psychologists think about participants. *Psychology of Women Quarterly*, *43*(4), 445–456. https://doi.org/10.1177/0361684319866430.

Rad, M. S., Martingano, A. J., & Ginges, J. (2018). Toward a psychology of homo sapiens: Making psychological science more representative of the human population. *Proceedings of the National Academy of Sciences*, *115*(45), 11401–11405. https://doi.org/10.1073/pnas.1721165115.

Reichel, P. L. (1988). Southern slave patrols as a transitional police type. *American Journal of Police*, *7*(2), 51–77.

Richeson, J. A., & Shelton, J. N. (2007). Negotiating interracial interactions: Costs, consequences, and possibilities. *Current Directions in Psychological Science*, *16*(6), 316–320. https://doi.org/10.1111/j.1467-8721.2007.00528.x.

Roberts, S. O., Bareket-Shavit, C., Dollins, F. A., Goldie, P. D., & Mortenson, E. (2020). Racial inequality in psychological research: Trends of the past and recommendations for the future. *Perspectives on Psychological Science*, *15*(6), 1295–1309. https://doi.org/10.1177/1745691620927709.

Rolin, K. (2009). Standpoint theory as a methodology for the study of power relations. *Hypatia*, *24*(4), 218–226. https://doi.org/10.1111/j.1527-2001.2009.01070.x.

Schultz, J. R., Gaither, S. E., Urry, H. L., & Maddox, K. B. (2015). Reframing anxiety to encourage interracial interactions. *Translational Issues in Psychological Science*, *1*(4), 392–400. https://doi.org/10.1037/tps0000048.

Shelton, J. N. (2000). A reconceptualization of how we study issues of racial prejudice. *Personality and Social Psychology Review*, *4*(4), 374–390. https://doi.org/10.1207/s15327957pspr0404_6.

Shelton, J. N. (2003). Interpersonal concerns in social encounters between majority and minority group members. *Group Processes & Intergroup Relations*, *6*(2), 171–185. https://doi.org/10.1177/1368430203006002003.

Thalmayer, A. G., Toscanelli, C., & Arnett, J. J. (2020). The neglected 95% revisited: Is American psychology becoming less American? *American Psychologist*. *76*(1), 116–129. https://doi.org/10.1037/amp0000622.

Vorauer, J. D. (2006). An information search model of evaluative concerns in intergroup interaction. *Psychological Review*, *113*(4), 862–886. https://doi.org/10.1037/0033-295X.113.4.862.

Warner, L. R. (2008). A best practices guide to intersectional approaches in psychological research. *Sex Roles*, *59*, 454–463. https://doi.org/10.1007/s11199-008-9504-5.

The internal validity obsession

Gregory Mitchell[a] and Philip E. Tetlock[b]

[a]University of Virginia, School of Law, Charlottesville, VA 22903, USA and [b]Psychology Department and Wharton School, University of Pennsylvania, Solomon Labs, Philadelphia, PA 19104, USA.
greg.mitchell@law.virginia.edu;
https://www.law.virginia.edu/faculty/profile/pgm6u/1191856
tetlock@wharton.upenn.edu; https://www.sas.upenn.edu/tetlock/bio

doi:10.1017/S0140525X21000637, e83

Abstract

Until social psychology devotes as much attention to construct and external validity as it does to internal validity, the field will continue to produce theories that fail to replicate in the field and cannot be used to meliorate social problems.

The target article joins a long line of compelling critiques of social psychology methodology. We suspect the latest critique, like its predecessors, will have little effect on how social psychologists study discrimination. A design ethos of "experimental realism" that relies on engaging but manufactured social settings (Aronson & Carlsmith, 1968) makes gathering data much easier than an approach that demands fidelity to real-world contingencies. The retort to critics is that the goal is to find theories that generalize and experimental control is essential to theory development (e.g., Banaji & Crowder, 1989).

Unfortunately, social psychologists rarely examine whether their theories do, in fact, generalize; when they do, the results are not pretty (Mitchell, 2012). Nor do social psychology journals demand much evidence that experimental constructions actually measure or manipulate the hypothesized processes of interest (Chester & Lasko, 2021), which helps explain why more than 20 years after the racial-attitudes implicit association test was introduced, we still do not know what it actually measures (Schimmack, 2021). The career calculus is clear. With journals happy for authors to speculate about the real-world implications of a statistically significant correlation or mean difference observed using convenience samples under artificial conditions, why embark on the arduous task of establishing external and construct validity? Any possible confound in design will spell doom for publication, while obvious shortcomings in the sample and manipulations chosen to test what passes for a theory will merit only cursory mention in a concluding section on limitations of the study.

As long as social psychology journals exalt internal validity over all other forms of validity, we should not expect social psychology to produce any theories that can really explain, much less help meliorate, social problems. Making passage of reality checks essential to publication (e.g., requiring comparison of an online convenience sample to a sample of persons with experience in the domain of interest or requiring that a theory be tested on archival data and not only on materials constructed for an experiment) would move the field away from exalting effects that prove to be the product of a quirky design decision that ignored key features of the situations or persons of theoretical interest. Such reality checks would serve as a form of "consistency test" of the kind that mature sciences employ (Meehl, 1978), and making reality checks a condition for publication would encourage greater care in theory development, pushing theorists to spell out boundary conditions and necessary auxiliary assumptions to narrow the range of reality checks that must be passed for the theory to survive.

We can understand why an exasperated Bayesian observer might conclude that until reality checks become a required part of theory validation within the field, the default assumption should be the best base-rate guess: neither social psychological theories nor effects will generalize. To those who worry that this default assumption would protect an oppressive status quo, we propose to locate the debate in signal detection framework. A false-negative error would be to dismiss a truly generalizable social psychological effect. A false-positive error would be to embrace an effect that proves to be a hot-house flower that wilts fast in the wild. We see the latter error as vastly more common today – hence our sympathy for the exasperated Bayesian. Our view is that it is better – for both the science and society – to require investigators to test the practical utility of their ideas using rigorous evaluation methods than to give politicians or consultants open-ended scientific license to invent popular or profitable interventions that they hope will work but that they never intend to subject to rigorous evaluation (see, e.g., Paluck, Porat, Clark, & Green, 2021).

Take the case of implicit bias. To our knowledge, no implicit bias training program implemented by a police department or other organization has ever been shown to have net behavioral benefits or to be justified under any cost–benefit analysis, yet countless dollars and work hours are being spent on such programs rather than other programs that might prove more effective. Certainly the belief that implicit bias explains many group disparities is widespread, and that belief may well have positive political consequences for some groups and may even reduce discrimination through increased sensitivity to its occurrence, but that belief continues to exist despite, not because of, social psychological research on the predictive (in)validity of measures of implicit bias. If the goal of social psychology is to create an ideology, rather than a science of social behavior, then it appears to have succeeded in the short term, but we suspect that success will erode its long-term credibility and its ability to provide long-term solutions to social problems.

Financial support. This research received no specific grant from any funding agency, commercial, or not-for-profit sectors.

Conflict of interest. None.

References

Aronson, E. R., & Carlsmith, J. M. (1968). Experimentation in social psychology. In G. Lindzey & E. Aronson (Eds.), *The handbook of social psychology* (Vol. 2, pp. 1–79). Reading, MA: Addison-Wesley.

Banaji, M. R. & Crowder, R. C. (1989). The bankruptcy of everyday memory. *American Psychologist*, *44*, 1185–1193. https://doi.org/10.1037/0003-066X.44.9.1185

Chester, D. S., & Lasko, E. M. (2021). Construct validation of experimental manipulations in social psychology: Current practices and recommendations for the future. *Perspectives on Psychological Science*, *16*, 377–395. https://doi.org/10.1177/1745691620950684

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, *46*, 806–834. https://doi.org/10.1016/j.appsy.2004.02.001

Mitchell, G. (2012). Revisiting truth or triviality: The external validity of research in the psychological laboratory. *Perspectives on Psychological Science*, *7*, 109–117.

Paluck, E. L., Porat, R., Clark, C. S., & Green, D. P. (2021). Prejudice reduction: Progress and challenges. *Annual Review of Psychology*, *72*, 533–560. https://doi.org/10.1146/annurev-psych-071620-030619

Schimmack, U. (2021). The implicit association test: A method in search of a construct. *Perspectives on Psychological Science*, *16*, 396–414. https://doi.org/10.1177/1745691619863798

External validity of social psychological experiments is a concern, but these models are useful

Youri L. Mora[a], Olivier Klein[a], Christophe Leys[a] and Annique Smeding[b]

[a]Center for Social and Cultural Psychology, Université libre de Bruxelles, CP 122, 50-1050 Bruxelles, Belgium and [b]Univ. Savoie Mont Blanc, LIP/PC2S, 73011 Chambéry cedex, France.
Youri.mora@ulb.be
Olivier.klein@ulb.be
Christophe.leys@ulb.be
Annique.Smeding@univ-smb.fr
https://cescup.ulb.be/member/youri-mora/
https://cescup.ulb.be/member/olivier-klein/
https://cescup.ulb.be/member/christophe-leys/
http://smeding.free.fr/index.php/Main/HomePage

doi:10.1017/S0140525X21000753, e84

Abstract

We agree that external validity of social psychological experiments is a concern, we disagree these models are useless. Experiments, reconsidered from a situated cognition perspective and non-linearly combined with other methods (qualitative and simulations) allow grasping decision dynamics beyond bias outcomes. Dynamic (vs. discrete) insights regarding these processes are key to understand missing forces and bias in real-world social groups.

In this commentary, we aim at extending Cesario's critique on the "use of experimental social psychology to explain real-world group disparities." While we agree that focusing on average bias may impair external validity in significant ways, we disagree with the lesson Cesario draws from the use of experimental paradigms: Using such paradigms should be a tool of last resort to explain real-world disparities. Indeed, situated cognition experiments have already demonstrated their usefulness in shedding light upon the decision dynamics involved in bias, and not only on the *presence* of bias, which is Cesario's focus. We will develop why drawing Cesario's lesson would amount to throwing the baby out with the bathwater and will propose alternative solutions to his.

First, experimental paradigms in social psychology serve as models of the real world and as such (a) are by definition a simplification (parsimony in modelling); but (b) are still useful to understand the psychological processes (Smaldino, 2017) that drive behaviour in the "real world." This concern is not new and is aptly illustrated by George Box's aphorism (1979, p. 202) "All models are wrong, but some are useful." What Cesario is essentially saying is that current models in experimental social psychology are wrong and not even useful to explain real-world disparities. The three flaws accurately identified by Cesario are actually three types of missing variables – moderators – whose absence is involved in deterioration of external validity. One can reformulate this criticism as suggesting that the studied effect sizes are smaller than the smallest effect size of interest when more ecological variables – absent in most experimental models – are taken into account. Appraising the problem through this lens leads us to disagree with Cesario's idea that the "research is *fundamentally* flawed." Still, the question remains: How to design experiments that yield more robust and meaningful effect sizes while accounting for these missing variables and which are applicable to real-life situations? The three identified flaws raise questions to which answers can help incrementally elaborate models that include crucial moderators.

Second, and related, we acknowledge that demonstrating average bias in "the general population" (often undergraduate, non-expert psychology students) with indirect measures like the implicit association test (IAT) may fall short in accounting for real-world group disparities. Nonetheless, using such paradigms to understand differential processes in real-world social groups has proven valuable. We were surprised that Cesario's section on implicit bias and science, technology, engineering, and mathematics (STEM)-related IAT did not refer to lines of research grounded in situated social cognition (Smith & Semin, 2007) or Freeman's work on social perception (Freeman, 2014; Freeman, Pauker, & Sanchez, 2016). For instance, using a mouse-tracking adapted gender-math IAT, completed by female and male STEM and non-STEM majors, research (Smeding, Quinton, Lauer, Barca, & Pezzulo, 2016) has shown meaningful group differences in decision-making dynamics (i.e., attractions) and their early emergence in time (around 300 ms). Millisecond differences – or deviations in mouse trajectories while decision-making is unfolding – may thus represent one of those real-world forces that characterise real-world group disparities. These can be measured with experimental paradigms. Also, Cesario frames the interest about bias in decision-making in the minds of "gatekeepers" discriminating against ambiguous candidates. However, candidates themselves will be the first depleting link in the decision-making chain if they are biased by held stereotypes (e.g., Shapiro and Williams, 2012) or because of imperfect inferences drawn from observed regularities (Kutzner & Fiedler, 2017); leading attrition to occur way before formal selection by potential gatekeepers. By comparing engineering and humanities female students, Study 3 in Smeding et al. (2016) has shown that self-congruency trumps the role of stereotype-congruency in a "Math versus Language" IAT. Self-congruency would here be categorised by Cesario as a "missing force." This moderator could not have been identified with a convenience sample. However, it can still be studied through an IAT paradigm, reconsidered from a situated social cognition approach. To explain underrepresentation of women in STEM majors, "men and women differ [ing] in their interest" (sect. 1, para. 6) can be considered as a mere demographic difference across groups, discarding the role of decision bias. However, the key role of self-congruency shows that the underrepresentation phenomenon can still be explained by decision bias (in the minds of candidates), which an IAT can meaningfully investigate while providing leads to implement

https://doi.org/10.1017/S0140525X21000017 Published online by Cambridge University Press

change (e.g., impact on self-domain related associations). Still, such investigation requires participants who are involved in the actual field of application and not convenience samples, as called by Cesario. Besides, mouse-tracking-based paradigms, anchored in dynamical systems theory (which has nurtured many real-life applications spanning from physics to cognitive science, Krpan, 2017) represent a user-friendly tool (Rivollier, Quinton, Gonthier, & Smeding, 2020) that allows understanding (a) continuity and (nonlinear) competition in decision-making (beyond discrete judgments) and (b) the influence of social category triggers specifically when people are ambiguous on a relevant real-world characteristic (Freeman, 2014; Freeman et al., 2016).

Finally, experiments – as one of the methods available to psychologists – in combination not solely with in-depth (qualitative) field research (as suggested by Cesario), but also computational modelling have the potential to provide insight into real-world human behaviour, including group disparities. In Smeding et al. (2016), results for simulated social groups and real-world social groups were compared. While the former provided proof of concept regarding the (hypothesised) psychological processes, the latter sustained their real-world validity. Both contributed to a finer-grained understanding of sex differences in STEM engagement which, admittedly, seemed to be less related to average stereotypic bias than to differential associations related to the self. But an experimental paradigm like the IAT happened to be of paramount importance in such findings. Mixed-methods (including qualitative, experimental, correlational, observational, computational, but also emerging real-world data-driven approaches such as machine learning) would all greatly benefit the study of bias and group disparities in social psychology. However, their use in a research programme is certainly nonlinear and more dynamic than the fixed sequence depicted by Cesario in his suggested new/rehashed approach.

Financial support. This work was supported by the Fund for Scientific Research (F.R.S.–FNRS) under Grant No. 40000042.

Conflict of interest. The authors declare no conflict of interest.

References

Box, G. E. P. (1979). Robustness in the strategy of scientific model building. *Robustness in Statistics*, 201–236. https://doi.org/10.1016/B978-0-12-438150-6.50018-2.

Freeman, J. B. (2014). Abrupt category shifts during real-time person perception. *Psychonomic Bulletin & Review 21*:85–92. https://doi.org/10.3758/s13423-013-0470-8.

Freeman, J., Pauker, K., & Sanchez, D. (2016). A perceptual pathway to bias. *Psychological Science 27*(4):502–517. https://doi.org/10.1177/0956797615627418.

Krpan, D. (2017). Behavioral priming 2.0: Enter a dynamical systems perspective. *Frontiers in Psychology 8*:1204. https://doi.org/10.3389/fpsyg.2017.01204.

Kutzner, F., & Fiedler, K. (2017). Stereotypes as pseudocontingencies. *European Review of Social Psychology 28*(1):1–49. https://doi.org/10.1080/10463283.2016.1260238.

Rivollier, G., Quinton, J.-C., Gonthier, C., & Smeding, A. (2020). Looking with the (computer) mouse: How to unveil problem-solving strategies in matrix reasoning without eye-tracking. *Behavior Research Methods*, *53*, 1081–1096. https://doi.org/10.3758/s13428-020-01484-3.

Shapiro, J. R., & Williams, A. M. (2012). The role of stereotype threats in undermining girls' and women's performance and interest in STEM fields. *Sex Roles 66*(3-4):175–183. https://doi.org/10.1007/s11199-011-0051-0.

Smaldino, P. E. (2017). Models are stupid, and we need more of them. In R. R. Vallacher, S. J. Read, & A. Nowak (Eds.), *Computational social psychology* (1st ed., pp. 311–331). Routledge. https://doi.org/10.4324/9781315173726-14.

Smeding, A., Quinton, J.-C., Lauer, K., Barca, L., & Pezzulo, G. (2016). Tracking and simulating dynamics of implicit stereotypes: A situated social cognition perspective. *Journal of Personality and Social Psychology 111*(6):817–834. https://doi.org/10.1037/pspa0000063.

Smith, E. R., & Semin, G. R. (2007). Situated social cognition. *Current Directions in Psychological Science 16*(3):132–135. https://doi.org/10.1111/j.1467-8721.2007.00490.x.

Controlled lab experiments are one of many useful scientific methods to investigate bias

Jason A. Okonofua

Department of Psychology, University of California, Berkeley, CA 94704, USA.
psychadmin@berkeley.edu

doi:10.1017/S0140525X21000650, e85

Abstract

Ecological validity is key in science and laboratory experiments alone cannot fully explain complex real-world phenomena. Yet the three flaws Cesario proposes do not characterize the field and are not "methodological trickery," (sect. 5, para. 5) designed to intentionally mislead practitioners. In school discipline alone, these alleged flaws are indeed addressed and laboratory experimentation has contributed to mitigation of a real-world problem.

Cesario's misrepresentation of the research designs of past studies, and overextension of his critique, risk irony given the topic of the article.

Missing contingencies flaw: addressed. Cesario criticizes experimental studies on bias by claiming that they use "novice or experimental participants" (e.g., undergrads) who are untrained decision-makers. Yet the experiments cited (Jarvis & Okonofua, 2020; Okonofua & Eberhardt, 2015) and other similar experiments (Okonofua, Paunesku, & Walton, 2016) have exclusively sampled hundreds of practicing K-12 teachers and principals. For example, Okonofua and Eberhardt (2015, Study 2) "recruited 204 K-12 teachers" (p. 4). In Table 3, Okonofua, Perez, and Darling-Hammond (2020) show how their sample of 243 teachers is overwhelmingly similar in demographic representation to the national K-12 teacher workforce.

Cesario also describes the information provided to study participants about student misbehavior as "impoverished descriptions of real teacher–child experiences" (sect. 4.3, para. 3). This claim also lacks factual merit. In the publications, the researchers describe specifically why the stimuli are representative descriptions of actual real teacher–child experiences. First, the stimuli presented describe the most common student misbehavior real teachers face (Losen & Martinez, 2013). Okonofua and Eberhardt (2015) write, "Minor infractions (e.g., for insubordination or class disruption) are the most frequently reported reasons for referring students to the principal's office" (p. 2). Second, the descriptions of the student misbehavior used in the study are taken directly from actual office-referral forms – using the precise words of a real K-12 teacher who referred a real student to a real principal's office for actual discipline. Okonofua and Eberhardt (2015) state: "[Teachers] then viewed a school record – adapted from actual office-referral records we collected from a public middle school in California" (p. 2). Rather than impoverished, participants read the exact same information that is presented in the real world. Third, the cited research (Jarvis & Okonofua, 2020) asks in-service principals – real-world decision-makers – to make discipline decisions based on this information.

Missing information flaw: addressed. Cesario claims that we removed "important information that real decision-makers could

use such as a child's history of behavior in the classroom" (sect. 4.3, para. 3). Ironically, the purpose of the experiments was specifically to examine how the history of a child's misbehavior influences educators' perceptions of the child and disciplinary decisions. As predicted, providing this history only increased bias; it in no way diminished it. This theoretical emphasis is present not only in the first two words of the publication's title "Two Strikes," but was deliberately embedded in the repeated-measures design that randomly counterbalanced the order of incidents in the cited experiments (Jarvis & Okonofua, 2020; Okonofua & Eberhardt, 2015).

Missing forces flaw: discussed at length. Cesario claims that the researchers expect "children who differ in myriad important ways should behave identically" (sect. 4.3, para. 5). This is also false. We do not claim that children will *always* behave identically; to the contrary, our theory is designed to understand specifically how real differences in student behavior and disciplinary outcomes arise (Okonofua et al., 2016). In this article, we go to lengths (+1,527 words) to describe what might lead children from different backgrounds to come to behave in different ways. Nevertheless, group differences in student misbehavior cannot fully account for racial disparities in discipline. And one need not take a psychologist's word for it. Education researchers' review of the latest education research conclude that

> Although low-income and minority students experience suspensions and expulsions at higher rates than their peers, these differences cannot be solely attributed to socioeconomic status or increased misbehavior. Instead, school and classroom occurrences that result from the policies, practices, and perspectives of teachers and principals appear to play an important role in explaining the disparities (Welsh & Little, 2018)

Using a tightly controlled experimental paradigm, in Okonofua and Eberhardt (2015), we show that even in cases where Black and White children do, indeed, behave the same, teachers do not treat them the same. In fact, differences in teacher treatment may be one factor (of many) that could lead to differences in student behavior down the road. Thus, our theory points to the self-fulfilling consequences of tying individual Black children to group-based stereotypes.

In the end, Cesario's article is as myopic and overextended as it accuses the field of bias research to be. This is manifestly apparent in its neglect of intervention field experiments – randomized placebo-controlled studies that draw directly on the insights and theory developed through laboratory experimentation and then uses these to reduce real-world discipline problems. The criticized Okonofua et al. (2016) publication spends more than 2,000 words reviewing such studies. It is by understanding how bias and apprehensions about bias can undermine teacher–student relationships – through laboratory experiments and basic theory – that these studies find ways to improve trajectories and outcomes. For example, Okonofua et al. (2016) show that the same experimental paradigm can be used to determine if an "empathic-mindset," a treatment to prioritize valuing students' perspectives when they misbehave, can reduce the likelihood a teacher will label a hypothetical Black student who misbehaves as a troublemaker (also see Okonofua et al., 2020). They then use this "empathic-mindset" approach in a field experiment with teachers who serve 1,682 actual students, which cut actual suspension rates over the academic year by 4.8 percentage points (also see Borman, Rozek, Pyne, & Hanselman, 2019; Goyer et al., 2019; Yeager et al., 2014).

The author's argument rests on a series of basic factual errors in describing controlled lab experiments on school discipline. The article does not acknowledge the contribution of controlled lab experiments to field experiments that have, in fact, dramatically reduced discipline in the real world. Instead of advancing theory or methodology, this article concludes by making a moral claim – that it is acceptable to judge individuals based on assumptions about social groups – that is contrary to public consensus and law, as though it were a scientific claim suitable for a science journal.

Acknowledgement. The author thanks Shoshana Jarvis, Gregory Walton, and Jennifer Eberhardt.

Financial support. This research received no specific grant from a funding agency, commercial, or not-for-profit sectors.

Conflict of interest. None.

References

Borman, G. D., Rozek, C. S., Pyne, J., & Hanselman, P. (2019). Reappraising academic and social adversity improves middle school students' academic achievement, behavior, and well-being. *Proceedings of the National Academy of Sciences*, *116*(33), 16286–16291.

Goyer, J. P., Cohen, G. L., Cook, J. E., Master, A., Apfel, N., Lee, W., … Walton, G. M. (2019). Targeted identity-safety interventions cause lasting reductions in discipline citations among negatively stereotyped boys. *Journal of Personality and Social Psychology*, *117*(2), 229.

Jarvis, S. N., & Okonofua, J. A. (2020). School deferred: When bias affects school leaders. *Social Psychological and Personality Science*, *11*(4), 492–498.

Losen, D. J., & Martinez, T. E. (2013). Out of school and off track: The overuse of suspensions in American middle and high schools.

Okonofua, J. A., & Eberhardt, J. L. (2015). Two strikes: Race and the disciplining of young students. *Psychological Science*, *26*(5), 617–624.

Okonofua, J. A., Paunesku, D., & Walton, G. M. (2016). Brief intervention to encourage empathic discipline cuts suspension rates in half among adolescents. *Proceedings of the National Academy of Sciences*, *113*(19), 5221–5226.

Okonofua, J. A., Perez, A. D., & Darling-Hammond, S. (2020). When policy and psychology meet: Mitigating the consequences of bias in schools. *Science Advances*, *6*(42), eaba9479.

Welsh, R. O., & Little, S. (2018). The school discipline dilemma: A comprehensive review of disparities and alternative approaches. *Review of Educational Research*, *88*(5), 752–794.

Yeager, D. S., Purdie-Vaughns, V., Garcia, J., Apfel, N., Brzustoski, P., Master, A., … Cohen, G. L. (2014). Breaking the cycle of mistrust: Wise interventions to provide critical feedback across the racial divide. *Journal of Experimental Psychology: General*, *143*(2), 804.

Culturally fluent real-world disparities can blind us to bias: Experiments using a cultural lens can help

Daphna Oyserman and Amabel Youngbin Jeon

Mind and Society Center, University of Southern California, Los Angeles, CA 90089, USA.
oyserman@usc.edu, youngbin@usc.edu
https://dornsife.usc.edu/daphna-oyserman, https://dornsife.usc.edu/mindandsociety/currentmembers/

doi:10.1017/S0140525X21000819, e86

Abstract

Culture provides people with rich, detailed, implicit, and explicit knowledge about associations (what goes together) and contingencies (how situations are likely to unfold). These

culture-based expectations allow people to get through their days without much systematic reasoning. Experimental designs that unpack these situated effects of culture on thinking, feeling, and doing can advance bias research and direct policy and intervention.

Cesario questions the role of experimental social-psychological bias research, suggesting that experiment-based bias research focuses only on social category membership, missing the context-specific interactions and features of persons that dominate decision-making. We agree with much of Cesario's concern, though not necessarily his conclusions. Here, we elaborate on our disagreement and suggest a particular path forward – experiments highlighting the roles cultural fluency and disfluency play in both bias emergence/maintenance and as obstacles to bias correction and anti-racism.

Social psychologists assume that people respond to situations as they construe them. Hence, social-psychologically grounded experiments focus on illuminating and understanding these construals. Experiments allow for clear tests of how construal affects what people do by maximizing researcher certainty about what people have on their minds in a situation. However, experiments do not randomly draw situations or behaviors from the population of all the situations/behaviors occurring in the real world. Instead, experiments set up a particular situation which people are likely to understand in a particular way. They test whether people put in that situation respond differently from people put in a psychologically distinct one. Lab-based experiments can test whether a specific construal process *could* be occurring by setting up situations in which people are likely to reach the same construal. They cannot test whether these construal processes *typically* occur or the relative *size of the effect* of tested construal processes outside the lab.

Cesario's target paper highlights these as limits to applying lab-based bias research to policy and intervention. We agree but disagree with the conclusion that experiments cannot be helpful. What is missing from the experiments Cesario critiques is a theoretical framework bridging to the real world.

We propose culture-as-situated-cognition theory as that bridge. Culture-as-situated-cognition theory (Oyserman, 2015) is a social-psychological theory of what culture is for and how it works. It predicts that living in a society yields cultural expertise in the form of culture-based knowledge residing in memory as associative knowledge networks. People automatically use that subset of their available culture-based knowledge accessible in the moment of judgment to make implicit predictions about what will happen next. When observations (e.g., a mournful obituary) seem to match culture-based expectations, they preserve people's sense that the world is as expected (so no thinking is needed), preserving cognitive resources (Oyserman & Yan, 2018; Oyserman, Novin, Flinkenflögel, & Krabbendam, 2014). After experiencing culturally fluent situations, people are more likely to accept the world-as-it-is and consequently see cultural groups as having more permanent, essential differences (Lin, Arieli, & Oyserman, 2019).

In contrast, people engage more carefully and process more deeply when their observations mismatch culture-based expectations. Mismatches (e.g., a delighted obituary) yield a metacognitive experience of disfluency, which signals that something is wrong without clarifying what precisely is wrong (Oyserman et al., 2014). Cultural (dis)fluency has consequences. People are less likely to use rule-based reasoning after experiencing culturally fluent rather than disfluent cues (Mourey, Lam, & Oyserman, 2015).

Experiments using a culture-as-situated-cognition approach can allow researchers to make progress in two ways. First, they pinpoint the easy-to-process features of the situation that match people's expectations. Second, they illuminate the consequence of experiencing expectation-observation mismatches (Oyserman, 2011, 2017). Each is a place that researchers should look for possible bias effects. By manipulating cultural fluency and studying what happens in culturally disfluent situations, researchers can unpack how cultural fluency works to shape and maintain bias and why correction and anti-bias are so non-obvious.

In the case of race-based stereotypical responses, people's culture-based associative knowledge networks include representations of how situations involving people from specific groups will likely unfold. These automatic predictions include appraisals (e.g., competence and trustworthiness), content-specific beliefs (stereotypes), emotions (prejudices), and behavioral responses (discriminatory tendencies, Dovidio & Fiske, 2012). These culture-based expectations shape how people construe their immediate situation. The reverse is also true. Group-based disparities rooted in discriminatory legislation/policies (e.g., red-lining and segregation) and differential resource access and control can create culture-based stereotypes about group features. People experience these disparities as culturally fluent group-based features people and use them to automatically predict how interactions will unfold when they anticipate interacting with people they expect to be from these groups. These culture-based expectations shape what people are likely to pay attention to in their interactions and whether they will stick to gut-based processing even if rule-based processing is needed. Moreover, even if people notice a mismatch between their culture-based expectation and the situation, people are unlikely to infer that bias is the problem. That is because mismatch only works to alert people that something is wrong, not what that might be (Oyserman, 2019). Alertness increases vigilance and suspicion but does not pinpoint what the problem is; bias itself is not automatically revealed or changed. Change occurs only with targeted intervention.

As Cesario notes, in more naturalistic settings, decision-makers may attribute their actions to features of the situation or interaction. They see the specifics. This makes it difficult to conclude that race-based biases play a role. A culture-as-situated-cognition perspective highlights that the taken-for-granted version of reality entails culture-based biased expectations. Culture-based expectations matter even though people are unlikely to notice them. Experiments are critical in shedding light on the possibility of bias because bias hides in culturally fluent blind spots. Only experiments can document that a bias *might* matter and *how* it might matter. Intervention and policy researchers need lab-based experiments that illuminate what is culturally fluent when people interact within and across divides, which features of situations preserve and which disrupt cultural fluency, and with what consequences.

Other designs are needed to address the questions of *when*, *how often*, and *how much* bias matters. Researchers can use ecological-momentary assessment to learn *when* (Newman & Stone, 2019) and diary studies to learn *how often* (Bolger, Davis, & Rafaeli, 2003). They can meld these with daily reconstruction methods to learn *how much* bias matters in real-world situations (Kahneman, Krueger, Schkade, Schwarz, & Stone, 2004). To gain an estimate of effect sizes, they may turn to simulations using each of the inputs gained from the prior methods.

https://doi.org/10.1017/S0140525X21000017 Published online by Cambridge University Press

Of course, social-psychological experiments cannot encapsulate every step of this process; that is not their purpose. However, experimental studies offer a crucial step in understanding and tackling real-world disparities by shedding light on culturally fluent blind spots. Progress in understanding bias requires taking cultural fluency seriously.

Financial support. This research received no specific grant from any funding agency, commercial, or not-for-profit sectors

Conflict of interest. None.

References

Bolger, N., Davis, A., & Rafaeli, E. (2003). Diary methods: Capturing life as it is lived. *Annual Review of Psychology 54*(1):579–616.

Dovidio, J. F., & Fiske, S. T. (2012). Under the radar: How unexamined biases in decision-making processes in clinical interactions can contribute to health care disparities. *American Journal of Public Health 102*(5):945–952.

Kahneman, D., Krueger, A. B., Schkade, D. A., Schwarz, N., & Stone, A. A. (2004). A survey method for characterizing daily life experience: The day reconstruction method. *Science (New York, N.Y.) 306*(5702):1776–1780.

Lin, Y., Arieli, S., & Oyserman, D. (2019). Cultural fluency means all is okay, cultural disfluency implies otherwise. *Journal of Experimental Social Psychology 84*:103822.

Mourey, J. A., Lam, B. C., & Oyserman, D. (2015). Consequences of cultural fluency. *Social Cognition 33*(4):308–344.

Newman, D. B., & Stone, A. A. (2019). Understanding daily life with ecological momentary assessment. In F. Kardes, P. M. Herr, & N. Schwarz (Eds.), *Handbook of research methods in consumer psychology* (pp. 259–275). Routledge.

Oyserman, D. (2011). Culture as situated cognition: Cultural mindsets, cultural fluency, and meaning making. *European Review of Social Psychology 22*(1):164–214.

Oyserman, D. (2015). Culture as situated cognition. In R. Scott & S. Kosslyn (Eds.), *Emerging trends in The social And behavioral sciences: An interdisciplinary, searchable, And linkable resource*. Wiley & Sons, Inc. ISBN 978-1-118-90077-2.

Oyserman, D. (2017). Culture three ways: Culture and subcultures within countries. *Annual Review of Psychology 68*:435–463.

Oyserman, D. (2019). Cultural fluency, mindlessness, and gullibility. In R. Baumeister & J. Forgas (Eds.), *The social psychology of gullibility* (pp. 255–275). Routledge Press.

Oyserman, D., Novin, S., Flinkenflögel, N., & Krabbendam, L. (2014). Integrating culture-as-situated-cognition and neuroscience prediction models. *Culture and Brain 2*(1):1–26.

Oyserman, D., & Yan, V. X. (2018). Making meaning: A culture-as-situated cognition approach to the consequences of cultural fluency and disfluency. In S. Kitayama & D. Cohen (Eds.), *Handbook of cultural psychology* (pp. 536–565). Guilford.

Two thousand years after Archimedes, psychologist finds three topics that will simply not yield to the experimental method

B. Keith Payne[a] and Mahzarin R. Banaji[b]

[a]Department of Psychology & Neuroscience, University of North Carolina at Chapel Hill, Chapel Hill, NC 27516, USA and [b]Department of Psychology, Harvard University, Cambridge, MA 02138, USA.
payne@unc.edu;
https://bkpayne.web.unc.edu
mahzarin_banaji@harvard.edu,
https://www.people.fas.harvard.edu/~banaji/

doi:10.1017/S0140525X21000790, e87

Abstract

Cesario argues that experiments cannot illuminate real group disparities because they leave out factors that operate in ordinary life. But what Cesario calls flaws are, in fact, the point of the experimental method. Of all the topics in science, we have to wonder why racial discrimination would be uniquely unsuited for investigating with experiments. The argument to give up the most powerful scientific method to study one of the hardest problems we confront is laughable.

In his target article, Cesario argues that laboratory experiments cannot shed light on real group disparities because they leave out *information*, *forces*, and *contingencies* that operate in ordinary life. These so-called flaws motivate his recommendation to abandon the use of psychology experiments for understanding three topics related to racial discrimination. Even the least astute reader of this paper will ask the obvious question: Why are these three topics singled out as uniquely unsuitable for experimental treatment? Why would this question be raised about topics of study that came into being by the wizardry of a tribe called *social psychologists*, who in mid-twentieth century changed the world of science by boldly asserting that problems as complex as *obedience to authority*, *bystander non-intervention*, and *minimal group effects*, could be studied in the same way as atoms and cells. What in the name of Lewin, Heider, and Festinger does Cesario mean when he states that the experimental method is uniquely unsuited to the study of these three topics in social psychology? Why not add to his chosen topics for the garbage heap other equally complex problems such as climate change to be outside the bounds of the experimental method? Surely it's not easy to create the glaciers of the Uttarakhand in the lab for study so surely we should abandon all study of the effects of global warming! So, we must again reassert that the laziest reader of the target article will yawn out one question: Why is Cesario selecting the three topics from one subfield (social psychology) of one science (psychology), as uniquely unsuited for experimental treatment? We too wondered why.

In the rest of this comment, we do not engage with any of the specific areas of research Cesario selects, as that does not matter. Instead, we flatly state that if his thesis is to be taken at all seriously, we would need to reevaluate all of experimental psychology. In fact, if Cesario is to be taken seriously, it is not just three areas in psychology that should be abandoned, but the entire enterprise of physics, chemistry, and biology, the National Science Foundation, and BBS itself, that should each be abandoned and immediately.

Philosophers, mathematicians, and astronomers have observed regularities in the world for centuries. But discovery accelerated dramatically when natural philosophers began creating controlled conditions that abstracted away many details of ordinary experience in order to gain experimental control. The past 2000 years of the scientific method is the reason we boast of human progress, whether it be rockets, submarines, and airplanes; synthetic polymers, the cathode ray, or the periodic table; the discovery of antibodies, the sequencing of the genome, and vaccines that eradicate diseases such as smallpox and control viruses such as COVID.

Only when thought experiments based on intuition in the real world gave way to test tubes and Petri dishes did we have a chance to understand reality. Galileo told stories about dropping balls from the Tower of Pisa but he did his actual work on artificial equipment by rolling them down inclined planes at home. It was the only way to control the wind and slow the fall enough to measure with accuracy. Eventually, the physician's trial and error gave way to randomized trials. The randomized experiment, keeping everything

constant and varying a single variable, remains not just a good way, but the dominant way to establish causality with confidence.

True, ordinary intuition has always had trouble with the scientific method and understanding the reason for varying single variables. Even today, reactionary forces impart the message that the earth is flat and that a god made humans and placed them on earth. But those of us who are the beneficiaries of basic education and reason roll our eyes at these misguided views. When a youngster makes such a remark, we explain that to answer a question such as "will a feather and a rock fall at equal or unequal speed" we must start with the idea of a vacuum and answer the question in a counterintuitive way. Left to Cesario's argument (that we must drop Galileo's balls from the Tower of Pisa given its real-world allure), we would have little to show for all our centuries of science.

What Cesario calls flaws – missing information, missing forces, and missing contingencies – are, in fact, the point of the experimental method. Researchers working in this area do not claim that laboratory experiments capture the complexity of the real world, because the purpose of experiments is not to reinstate the real world in the lab. Experiments in this tradition look at one factor at a time, such as race, but also time pressure, anxiety, motivations to be unbiased, the identity of the subject, and the difference between police officers and civilians, just to name a few (for a review see Payne & Correll, 2020). For those doing this work, results of lab experiments are in constant conversation with other research, such as field experiments and observational studies of real-world disparities. For example, geographical patterning of race bias based on experimental tasks can be predicted by patterns of enslavement before the Civil War (Payne, Vuletich, & Brown-Iannuzzi, 2019). Patterns of implicit bias across countries can predict educational disparities in actual standardized tests (Nosek et al., 2009). And cities where residents more easily associate Black people with weapons on laboratory tasks have larger racial disparities on actual police use of force (Hehman, Flake, & Calanchini, 2018). To claim that experimentalists make inferences from experiments to everyday discrimination without doing the work of empirically integrating data at multiple levels of analysis is both naïve and factually simply wrong.

The observation that experiments sometimes lack realism is not new. Gergen (1973) argued that the findings of social psychology change with history and culture, casting doubt on whether laboratory experiments can produce insights that are general and cumulative. Neisser's (1978) call for the study of everyday memory lamented that memory research doesn't answer enough interesting or socially significant questions. The trade-off between realism and experimental control is well understood (Banaji & Crowder, 1989). But even these critics called for increased attention and integration, not simply abandoning the experimental method. That dubious innovation is new with Cesario's target article and it should be abandoned with haste unless the goal is to make a mockery of scientific psychology.

Financial support. This research received no specific grant from any funding agency, commercial, or not-for-profit sectors.

Conflict of interest. None.

References

Banaji, M. R., & Crowder, R. G. (1989). The bankruptcy of everyday memory. *American Psychologist 44*, 1185–1193.

Gergen, K. J. (1973). Social psychology as history. *Journal of Personality and Social Psychology 26*, 309.

Hehman, E., Flake, J. K., & Calanchini, J. (2018). Disproportionate use of lethal force in policing is associated with regional racial biases of residents. *Social Psychological and Personality Science 9*, 393–401.

Neisser, U. (1978). Memory: What are the important questions? In M. M. Gruneberg, E. E. Morris, & R. N. Sykes (Eds.), *Practical aspects of memory* (pp. 3–24). Academic Press.

Nosek, B. A., Smyth, F. L., Sriram, N., Lindner, N. M., Devos, T., Ayala, A., … Greenwald, A. G. (2009). National differences in gender–science stereotypes predict national sex differences in science and math achievement. *Proceedings of the National Academy of Sciences 106*, 10593–10597.

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

Payne, B. K., Vuletich, H. A., & Brown-Iannuzzi, J. L. (2019). Historical roots of implicit bias in slavery. *Proceedings of the National Academy of Sciences 116*, 11693–11698.

The call for ecological validity is right but missing perceptual idiosyncrasies is wrong

Jennie Qu-Lee and Emily Balcetis

Psychology Department, New York University, New York, NY 10003, USA.
jennie.qulee@nyu.edu
emilybalcetis@nyu.edu
https://sites.google.com/nyu.edu/nyu-spam-lab/

doi:10.1017/S0140525X21000807, e88

Abstract

Although psychology has long professed that perception predicts action, the strength of the evidence supporting the statement depends on the ecological validity of the technologies and paradigms used, particularly those that track eye movements, supporting Cesario's argument. While right to call for ecological validity, Cesario's model fails to account for individual differences in visual experience perceivers have when presented with the same stimulus.

What people see predicts what people do (Gibson, 1979). However, when researchers test connections between perception and action in contexts or when using paradigms that lack ecological validity, the utility of their conclusions is suspect. Aligning with Cesario's perspective, researchers across multiple fields that involve eye-tracking have highlighted discrepancies that emerge as a function of the presence – or rather, lack – of ecologically valid testing procedures. In the early 1900s perceivers wore contact lenses with attached pointers. In the 1930s, perceivers sat with their chin, forehead, and back of their head affixed to a metal frame attached to the desk. By the 1990s, perceivers wore bulky headgear with relatively large suspended cameras. These techniques thwart attempts to capture natural viewing experiences. Some developmental psychologists, aware of these limitations, created mobile eye-tracking, in which small cameras are affixed to baseball caps or eyeglasses that toddlers, children, and adults wear (Franchak, Kretch, Soska, & Adolph, 2011). Using this technology, researchers found that infants spent far less time looking at their mother's face during social interactions (Franchak et al., 2011; Jung, Zimmerman, & Pérez-Edgar, 2018), than previous research had concluded – research that used equipment requiring infants sit immobile at desk-mounted eye-trackers (Soska, Adolph, & Johnson, 2010).

When industrial research adopted ecologically valid eye-tracking technology, they discovered previous conclusions had been wrong too. When eye-tracking allowed pilots to engage with the simulation screen freely without imposing restrictions on head and body movements, researchers found that expert pilots in cockpits allocated attention in ways that gathered critical information during tactical operations (Li, Chiu, Kuo, & Wu, 2013; Pérez-Edgar, MacNeill, & Fu, 2020). Previously, the field had relied on restrictive eye-tracking technology with chinrests, and reported pilots failed to attend to necessary information (Sulzer & Skelton, 1976).

In our social cognition lab, we monitor eye movements using infrared sensors embedded into the frame of a typical-looking monitor, which records eye movements without awareness and as individuals freely move their heads and torsos within a 42° space. Under these naturalistic viewing experiences, we discovered that it was only among participants who frequently attended to an officer in a police–civilian altercation that pre-existing feelings of identification with police influenced legal decisions (Granot, Balcetis, Schneider, & Tyler, 2014). These data reconciled discrepancies between empirical studies and real trial data in the courts that have been inconclusive as to whether people punish outgroup members more harshly (Anwar, Bayer, & Hjalmarsson, 2012), more leniently (McGowen & King, 1982), or without bias (Mazzella & Feingold, 1994).

We also agree with Cesario that removing social context undermines ecological validity. Context offers a metaphorically thicker rather than thinner slice of information that informs understanding. Social context comes in many forms including the dimension of time. When we incorporated time by presenting dynamic rather than static visual scenes of the police aggression, we discovered individual differences in attention patterns that predicted why and when bias in police punishment decisions emerged (Granot et al., 2014).

However, we disagree with Cesario's conclusion that decision-makers including police and teachers respond without prejudice to the behaviors exhibited by individuals with whom they engage. In his "Missing Contingencies Flaw" tenet, Cesario argues that behaviors presented to decision-makers differ. This reflects an error of naïve realism. We argue people do respond with bias to the same stimulus because they do not perceive the same stimulus the same way in all cases. Ample evidence finds that individuals experience idiosyncratic perceptual experiences for at least two reasons. First, the demands on attention are greater than the attentional resources perceivers have available. As a result, attention is selective and directed (Broadbent, 1958; John, Bartlett, Shimokochi, & Kleinman, 1973); given that attention drives visual experience, differences in attention produce differences in perception (Mack & Rock, 1998). Moreover, differences in attention are systemic and vary as a function of characteristics of perceivers themselves. For instance, individuals attend to sources of threat, particularly when threats are members of a social outgroup (Koster, Crombez, Van Damme, Verschuere, & De Houwer, 2004). Indeed, White participants fixated longer on the civilian compared to an officer when viewing video evidence of both engaged in a physical altercation; though Asian participants did too, they showed significantly greater parity than did White participants (Sternisko, Granot, & Balcetis, 2017). Moreover, greater visual attention on the officer increased the severity of punishment leveraged against him. Although the stimulus participants responded to was the same, the manner in which individuals engaged attention varied systematically, resulting in differences in perceptual interpretation.

Second, because foveal view where details are encoded with great clarity constitutes only a small subset of the field of vision, most information that enters the visual system is processed through peripheral vision which is specialized for detecting movement, but little else (Fairchild, 2005). As a result, much of visual input is ambiguous and idiosyncratically interpreted. Indeed, even when presented with the same line drawing, individuals reached markedly different understandings of what they saw as a function of what they had previously been thinking about (Balcetis & Dunning, 2006). Moreover, when encouraged to reflect on their visual experience and reconstrue their percept in alternate ways, they could not. This suggests their initial biases affected encoding in such a way that visual cues relevant to reinterpretation were removed.

In other work, even when attention was experimentally held constant and participants attended to the same social target, prior attitudes biased perception of the attended stimulus (Granot et al., 2014). Even when focusing attention to the officer, perceivers differed in the degree to which they believed they saw the officer initiate physical contact, search the civilian, display a weapon, and pursue the civilian. These are objective, discrete, observable behaviors, but perceivers differed in the degree to which they believed they saw them happen even when attending to the same social target.

While we agree with Cesario there are limitations in extrapolating real-world consequences from experimental findings given that differences in ecological validity impact outcomes, it is an error to assume that all individuals perceive the same stimulus the same way.

Financial support. This research received no specific grant from any funding agency, commercial, or not-for-profit sectors.

Conflict of interest. None.

References

Anwar, S., Bayer, P., & Hjalmarsson, R. (2012). The impact of jury race in criminal trials. *The Quarterly Journal of Economics 127*(2):1017–1055.

Balcetis, E., & Dunning, D. (2006). See what you want to see: Motivational influences on visual perception. *Journal of Personality and Social Psychology 91*:612–625.

Broadbent, D. E. (1958). The selective nature of learning. In D. E. Broadbent (Ed.), *Perception and communication* (pp. 244–267). Pergamon Press.

Fairchild, M. D. (2005). *Color appearance models* (2nd Edn). John Wiley & Sons.

Franchak, J. M., Kretch, K. S., Soska, K. C., & Adolph, K. E. (2011). Head-mounted eye tracking: A new method to describe infant looking. *Child Development 82*(6):1738–1750.

Gibson, J. J. (1979). *The ecological approach to visual perception*. Houghton Mifflin.

Granot, Y., Balcetis, E., Schneider, K. E., & Tyler, T. R. (2014). Justice is not blind: Visual attention exaggerates effects of group identification on legal punishment. *Journal of Experimental Psychology: General 143*(6):2196–2208.

John, E. R., Bartlett, F., Shimokochi, M., & Kleinman, D. (1973). Neural readout from memory. *Journal of Neurophysiology 36*(5):893–924.

Jung, Y. J., Zimmerman, H. T., & Pérez-Edgar, K. (2018). A methodological case study with mobile eye-tracking of child interaction in a science museum. *TechTrends 62* (5):509–517.

Koster, E. H., Crombez, G., Van Damme, S., Verschuere, B., & De Houwer, J. (2004). Does imminent threat capture and hold attention?. *Emotion, 4*(3), 312–317.

Li, W. C., Chiu, F. C., Kuo, Y. S., & Wu, K. J. (2013). The investigation of visual attention and workload by experts and novices in the cockpit. In International Conference on Engineering Psychology and Cognitive Ergonomics (pp. 167–176). Springer.

Mack, A., & Rock, I. (1998). Inattentional blindness: Perception without attention. In R. Wright (Ed.), *Visual attention* (pp. 112–189). Oxford University Press.

Mazzella, R., & Feingold, A. (1994). The effects of physical attractiveness, race, socioeconomic status, and gender of defendants and victims on judgments of Mock Jurors: A meta-analysis 1. *Journal of Applied Social Psychology 24*(15):1315–1338.

McGowen, R., & King, G. D. (1982). Effects of authoritarian, anti-authoritarian, and egalitarian legal attitudes on mock juror and jury decisions. *Psychological Reports 51* (3_suppl):1067–1074.

Pérez-Edgar, K., MacNeill, L. A., & Fu, X. (2020). Navigating through the experienced environment: Insights from mobile eye tracking. *Current Directions in Psychological Science 29*(3):286–292.

https://doi.org/10.1017/S0140525X21000017 Published online by Cambridge University Press

Soska, K. C., Adolph, K. E., & Johnson, S. P. (2010). Systems in development: Motor skill acquisition facilitates three-dimensional object completion. *Developmental Psychology 46*(1):129.

Sternisko, A., Granot, Y., & Balcetis, E. (2017). One-sighted: How visual attention biases legal decision-making. In New Developments in Visual Attention Research (pp. 105–139). Nova Science Publishers.

Sulzer, R. L., & Skelton, G. E. (1976). Visual Attention of Private Pilots, the Proportion of Time Devoted to Outside the Cockpit. National Aviation Facilities Experimental Center, Atlantic City, NJ.

Developmental research assessing bias would benefit from naturalistic observation data

Jennifer L. Rennels and Kindy Insouvanh

Department of Psychology, University of Nevada, Las Vegas, Las Vegas, NV 89154-5030, USA.
Jennifer.rennels@unlv.edu; insouvan@unlv.nevada.edu
https://rebellab.sites.unlv.edu

doi:10.1017/S0140525X21000765, e89

Abstract

Cesario's critiques and suggestions for redesigning social psychology experiments echo Dahl's (2017) call for developmental researchers to use experimental and naturalistic methods in a complementary manner for understanding children's development. We provide examples of how naturalistic observations can rectify Cesario's *missing flaws* for developmental studies investigating children's social biases and help researchers derive theories they can then experimentally test.

Cesario identifies three broad concerns regarding how social psychologists design their experimental research examining individuals' displays of social biases. These include critiques regarding: (1) limited information about the targets, such as presenting targets that are similar on many social dimensions except for one attribute (*missing information flaw*); (2) oversight of factors other than social bias that might contribute to the outcome of interest (*missing forces flaw*); and (3) whether results translate to real-world scenarios and decision-making (*missing contingencies flaw*). These critiques are not limited to social psychology and are akin to Dahl's (2017) critiques of developmental research. When designing developmental studies, researchers control what participants see or experience and often what the possible responses are. Consequently, researchers make assumptions (what Dahl referred to as ecological commitments) about how children think, behave, and emotionally respond. They also postulate what experiences children have outside of the lab setting that contribute to developmental changes. Without actually observing individuals in their natural environment, researchers need to be particularly careful about whether their findings have ecological validity.

Cesario concludes the article by calling for a different approach to investigating social biases. The proposed methodology includes: (1) learning how decision-making ensues in real-world settings and what training/modeling occurs to support this process; (2) assessing inequities that members of particular social groups experience and what disparities beyond categorical membership perceivers consider in the decision-making process; and (3) using this information to create experimental studies. These suggestions echo Dahl's (2017) call to use experimental and naturalistic methods in a complementary manner for understanding children's development. For example, Rennels and Langlois (2014) compared 3- to 11-year-olds' explicit biases based on facial attractiveness, gender, and race and found that biases based on girls' facial attractiveness were the most robust. In this study, participants saw faces of two children who differed in attractiveness, gender, or race but had similar attributes otherwise. Their task was to assign positive and negative attributes to the children depicted. In the forced choice condition, participants had to choose one of the two children when assigning attributes. In the non-forced condition, participants could choose one of the two children, or both or neither child. Although the non-forced choice condition permitted more flexibility in how participants responded, the study provided no information regarding the targets other than appearance. This *missing information flaw* could be rectified by observing children in their natural environments where classmates' faces vary on more than one attribute and children have developed knowledge regarding the behavior of other children in the classroom. Bias could be assessed by documenting approach/avoidance behaviors and the positivity/negativity of interactions between children. If Rennels and Langlois' (2014) results generalize to natural environments, then compared to other classmates, children should be most likely to avoid and negatively interact with low attractive girls, and most likely to approach and positively interact with high attractive girls. It would also be important to assess whether teachers/staff model such differential behavior when interacting with children in the classroom who differ in attractiveness.

In terms of the *missing forces flaw*, a developmental example is the interpretation of children's gender biases and preferences for the same-gendered peers. It is well established that preschool and elementary children spend most of their time with the same-gendered peers who have similar interests in gender-typed activities (Martin et al., 2013). Yet this gender bias varies based on children's reinforcement of gender stereotypes – boys typically adhere more strictly to gender roles than girls (Katz & Walsh, 1991), potentially because socialization teaches individuals to value masculine activities more than feminine activities. For example, school-aged children preferred a girl who engaged in masculine activities as a potential classmate more than a boy who engaged in feminine activities (Braun & Davidson, 2017). Children's friendship and activity preferences, therefore, not only reflect gender similarity but also their endorsement of gender stereotypes and evaluation of masculine and feminine activities. Naturalistic observations could complement these experimental findings and reveal additional missing forces by documenting factors contributing to the quality and length of interactions between the same and mixed gender peers.

Applying the critique of the *missing contingencies flaw* to developmental research could provide insight regarding why children's displays of explicit biased attitudes do not consistently translate across situations into discriminatory behavior (Dunham & Degner, 2013). One experimental contingency often overlooked when examining children's biases is the extent to which their usual real-world scenarios might incentivize them to express or control their biases. For instance, when a researcher told 6- to 10-year-olds that other adults and children would see their responses to an explicit racial attitude questionnaire, those children showed less explicit bias than children who were told their responses would not be shared (Rutland, Cameron, Milne, & McGeorge, 2005). Thus, children can be externally motivated to

inhibit displays of explicit bias. With development, children's internal motivation to inhibit displays of bias becomes contingent upon their understanding of others' attitudes and emotions (i.e., theory of social mind [ToSM]; Abrams, Rutland, Pelletier, and Ferrell, 2009). High ToSM enables children to internalize their ingroup's social norms and inhibit displays of bias, whereas low ToSM requires external motivation to inhibit such displays (Fitzroy & Rutland, 2010). Conducting naturalistic observations of how students, teachers, and administrators respond to bias in conjunction with reviews of school policies could provide relevant data regarding what incentivizes individuals to display or inhibit bias.

Our recommendation is to use naturalistic observations to complement, validate, or negate experimental developmental findings and is not limited to studying individuals. As per the missing contingencies example, these recommendations should extend beyond assessments of individual level bias to consider participants' real-world settings. For example, institutional factors, such as a school's diversity, equity, and inclusion policies and actions, are typically not included in developmental explanatory models of biased behavior. Often, only the school's racial makeup is provided (e.g., McGlothlin and Killen, 2010). We encourage developmental researchers to use naturalistic settings to enhance our understanding of when, why, and how children display bias in real-world scenarios. Theories derived from such observations could then be experimentally tested (Dahl, 2017).

Financial support. Development of these ideas was supported by sabbatical leave the University of Nevada, Las Vegas awarded to Jennifer Rennels.

Conflict of interest. The authors declare no conflicts of interest.

References

Abrams, D., Rutland, A., Pelletier, J., & Ferrell, J. M. (2009). Children's group nous: Understanding and applying peer exclusion within and between groups. *Child Development*, *80*(1), 224–243. https://doi.org/10.1111/j.1467-8624.2008.01256.x.

Braun, S. S., & Davidson, A. J. (2017). Gender (non) conformity in middle childhood: A mixed methods approach to understanding gender-typed behavior, friendship, and peer preference. *Sex Roles*, *77*(1), 16–29. https://doi:10.1007/s11199-016-0693-z.

Dahl, A. (2017). Ecological commitments: Why developmental science needs naturalistic methods. *Child Development Perspectives*, *11*(2), 79–84. https://doi.org/10.1111/cdep.12217.

Dunham, Y., & Degner, J. (2013). From categories to exemplars (and back again). In M. R. Banaji & S.A. Gelman (Eds.), *Navigating the social world: What infants, children, and other species can teach us* (pp. 275–280). Oxford University Press.

Fitzroy, S., & Rutland, A. (2010). How social experience is related to children's intergroup attitudes. *European Journal of Social Psychology*, *40*(4), 679–693. https://doi.org/10.1002/ejsp.

Katz, P. A., & Walsh, P. V. (1991). Modification of children's gender-stereotyped behavior. *Child Development*, *62*(2), 338–351. https://doi.org/10.1111/j.1467-8624.1991.tb01535.x.

Martin, C. L., Kornienko, O., Schaefer, D. R., Hanish, L. D., Fabes, R. A., & Goble, P. (2013). The role of sex of peers and gender-typed activities in young children's peer affiliative networks: A longitudinal analysis of selection and influence. *Child Development*, *84*(3), 921–937. https://doi.org/10.1111/cdev.12032.

McGlothlin, H., & Killen, M. (2010). How social experience is related to children's intergroup attitudes. *European Journal of Social Psychology*, *40*(4), 625–634. https://doi.org/10.1002/ejsp.

Rennels, J. L., & Langlois, J. H. (2014). Children's attractiveness, gender, and race biases: A comparison of their strength and generality. *Child Development*, *85*(4), 1401–1418. https://doi.org/10.1111/cdev.12226.

Rutland, A., Cameron, L., Milne, A., & McGeorge, P. (2005). Social norms and self-presentation: Children's implicit and explicit intergroup attitudes. *Child Development*, *76*(2), 451–466. https://doi.org/10.1111/j.1467-8624.2005.00856.x.

The logic of challenging research into bias and social disparity

Regina Rini

Department of Philosophy, York University, Toronto, ON M3J 1P3, Canada.
rarini@yorku.ca
reginarini.net

doi:10.1017/S0140525X21000777, e90

Abstract

There are two problems with the logic of Cesario's argument for abandoning existing research on social bias. First, laboratory findings of decisional bias have social significance even if Cesario is right that the research strips away real-world context. Second, the argument makes overly skeptical demands of a research program seeking complex causal linkages between micro- and macro-scale phenomena.

Yes, our techniques for studying social disparities have some methodological weaknesses. But Cesario says something much stronger than this. Regarding two decades' research across many dozens of scholarly projects, he says in his abstract that "the current research tradition should be abandoned" (abstract). Now that's a conclusion! But iconoclasm is merited only if the argument is smashing. I will show two central flaws in Cesario's reasoning, either of which neutralizes his ambitious conclusion. Importantly, I will grant (for the sake of argument) Cesario's interpretive claims about the empirical literature. My two objections are instead about the *logic* of Cesario's argument; I will show that even if he is right about how to read these experiments, it is premature to recommend abandoning, or even drastically revising, the research tradition.

First, consider Cesario's claim that laboratory studies of bias strip away context from real-world decisions (the flaws of "missing information" and "missing contingencies," in Cesario's terms). Let's grant this for the sake of argument. How does it support the conclusion that the current research tradition should be abandoned? Because, Cesario says, such artificially barren laboratory decisions cannot predict real-world decisions.

But Cesario is wrong to assume that decision-prediction is the only socially relevant use of this research. He does concede that the research bears on the apparently anodyne question of "the function and process of storing and using categorical information" (sect. 5, para. 1). But there is something else which directly touches on the social questions that researchers take themselves to be addressing. It is this: These laboratory studies show that (at minimum) people tend to treat social categories like race and gender as *arbitrary-decision resolvers*. And that is an important fact to study.

To see the point, recall the fable of Buridan's ass. Faced with two piles of hay, each equally tempting, the donkey starves to death for lack of reason to resolve an arbitrary choice. Human agents have ways of avoiding this fate – we might flip a coin, or perhaps favor whatever is closest to our dominant hand. That's perfectly fine, so far. But our choice of arbitrary-decision resolver can have ethical implications.

Imagine you have two children and you have just won a sweepstakes that entitles them to random items from an expensive toy catalog. You are given a list of lot numbers and told to divide

them up between the kids. You have no idea *which* toy is represented by which lot number (that's the "fun" part, according to the toy manufacturer). If you had this information, of course, you might decide which child would like which toy more. But you don't. How should you resolve this arbitrary decision?

Here's a wrong answer: give 80% of the toys to one child, and 20% of the toys to the other child. This is the wrong answer because it displays inappropriate favoritism among your children. And it's no defense to insist that, because the choice was arbitrary, your decision resolution practice can be anything you want. Sometimes, how we choose to resolve an arbitrary decision reveals a great deal about the respect and care we have for the people affected. (Seriously, ask your kids.)

So, even if Cesario is right that these laboratory studies are arbitrary-decision contexts, that doesn't eliminate their social and political importance. It is important to know whether certain groups are implicitly treated as "less-than" even in arbitrary contexts. Ethicists have recently demonstrated how systemic derogation of a group constitutes disrespect even if it leads to no further consequences (Basu, 2019). Further, even when stereotypes are ostensibly supported by statistical group regularities (as Cesario suggests at times), this doesn't prevent individual decisions from being morally risky (Bolinger, 2020; Moss, 2018). All of which means this research tradition is valuable *even if* Cesario's criticisms are right.

My second objection to Cesario's logic concerns his apparent theory of how social scientists should synthesize reasoning across micro- and macroscopic causal phenomena. Here's what I mean. We have strong evidence of a micro-scale phenomenon: bias in lab conditions. We also have strong evidence of a macro-scale phenomenon: systemic outcome inequities in housing, employment, and policing. What we don't (yet) have is conclusive evidence of the micro-to-macro causal linkages between these two phenomena – though researchers are working on it (Mallon, 2021). Finding those linkages will take a long time, given that the causal system is enormously complicated. While that work is ongoing, the approach is especially vulnerable to skeptical challenges from alternative causal explanations. Cesario presents one: group differences. He suggest (in his "missing forces" argument) that social scientists must take more seriously the possibilities that Black citizens simply are more connected to violent crime, or that women simply are less qualified in science, technology, engineering, and mathematics (STEM) fields.

The problem is that Cesario overemphasizes the significance of these alternative theories. To see this point, consider the parallel to climate change skepticism. There we have another micro-scale phenomenon (thermal properties of carbon) and another macro-scale phenomenon (historical change in average global temperature), with complex and still not-fully-understood causal linkages between them. The skeptic presents an alternative explanation: natural epochal temperature cycles. The skeptic insists we cannot focus attention on carbon emissions until we have nailed down our micro-to-macro causal linkages and ruled out their alternative explanation.

Cesario isn't doing exactly this, but his argument is not too far off. If his point were simply that a complete science of social disparity will ultimately need to rule out group differences (just as complete climate science needs to rule out temperature cycles), then fair enough. But he goes far beyond this when he suggests the need to radically restructure, or even "abandon," the existing research paradigm. Research on social decision biases is only two or three decades old. Demanding airtight causal demonstration from it this early is comparable to judging the theory of anthropogenic climate change by the state of science in 1990.

Financial support. This research received no specific grant from any funding agency, commercial, or not-for-profit sectors.

Conflict of interest. None.

References

Basu, R. (2019). What we epistemically owe to each other. *Philosophical Studies*, *176*(4), 915–931.

Bolinger, R. J. (2020). The rational impermissibility of accepting (some) racial generalizations. *Synthese*, *197*(6), 2415–2431.

Mallon, R. (2021). Racial attitudes, accumulation mechanisms, and disparities. *Review of Philosophy and Psychology*, *12*, 953–975. https://doi.org/10.1007/s13164-020-00521-6.

Moss, S. (2018). Moral encroachment. *Proceedings of the Aristotelian Society*, *118*(2), 177–205.

The only thing that can stop bad causal inference is good causal inference

Julia M. Rohrer[a], Stefan C. Schmukle[a] and Richard McElreath[b]

[a]Department of Psychology, Leipzig University, D-04109 Leipzig, Germany and [b]Department of Human Behavior, Ecology, and Culture, Max Planck Institute for Evolutionary Anthropology, D-04103 Leipzig, Germany.
julia.rohrer@uni-leipzig.de
schmukle@uni-leipzig.de
richard_mcelreath@eva.mpg.de
www.juliarohrer.com
https://home.uni-leipzig.de/diffdiag/pppd/?page_id=101
https://xcelab.net/rm/

doi:10.1017/S0140525X21000789, e91

Abstract

In psychology, causal inference – both the transport from lab estimates to the real world and estimation on the basis of observational data – is often pursued in a casual manner. Underlying assumptions remain unarticulated; potential pitfalls are compiled in post-hoc lists of flaws. The field should move on to coherent frameworks of causal inference and generalizability that have been developed elsewhere.

Claims in psychology (and elsewhere) often rest on unarticulated, unconvincing assumptions, and we agree with much of Cesario's criticism. Yet the structure of his argument is symptomatic of the intuitive, implicit style of causal inference that contributes to careless conclusions. Outside of experiments, psychologists don't like explicit causal inference (Grosz, Rohrer, & Thoemmes, 2020); the "C-Word" (Hernán & Robins, 2010) is avoided, even if the discussion hinges on a causal interpretation. If instead an experiment was conducted, causal claims are accepted. But how the effect estimate from the experiment can be transferred to the rest of the world is left unarticulated, even if the discussion hinges on such transportability. This "inference by omission" can go wrong, and so psychologists compile lists of threats to validity, to which Cesario's "fatal flaws" could be appended. Unfortunately, lists of problems are not solutions.

Causal inference frameworks, such as the potential outcomes model (see Hernán & Robins, 2010, for a comprehensive

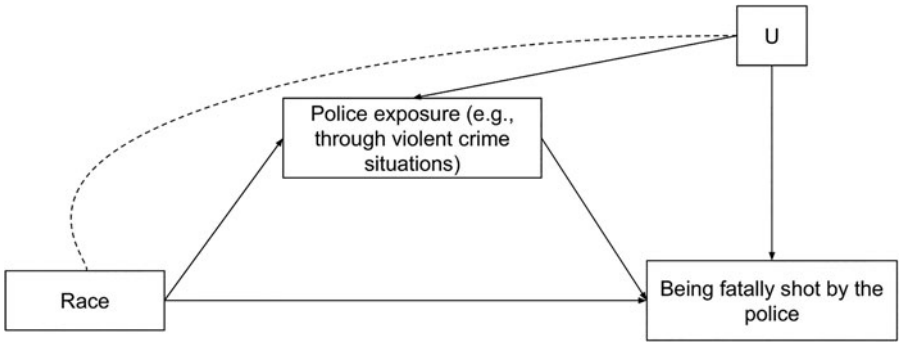

Figure 1 (Rohrer et al.). Mediational claim implicit in the notion that violent crime rates "account for" disparities in being fatally shot by the police. Police exposure is a collider variable between race and U, conditioning on it induces spurious associations between the two.

introduction; Little & Rubin, 2000) and graphical causal models (e.g., Pearl, Glymour, & Jewell, 2016; see also Rohrer, 2018, for an introduction for psychologists), provide rigorous formalization that aids in spelling out assumptions and deriving their implications. This explicitly causal lens can improve research design, analysis, and interpretation for experiments and non-experiments alike, so let us apply it to the question of decision-maker bias.

For the case of experiments, we can simplify Cesario's list of flaws. His major concern is effect modification: The effect of group membership depends on so-called effect modifiers (e.g., disambiguating information and decision-maker features). The distribution of effect modifiers in the experimental setting differs from the distribution in the setting which we want to make statements about. If the experimental setting holds the effect modifier constant at a value that does not occur in the target setting, transport of the estimate is impossible. But, if the experimental setting includes plausible values of the effect modifier, transport becomes possible under certain licensing assumptions (Pearl & Bareinboim, 2014). An understanding of these assumptions would help psychologists to systematically improve their studies for inferences about effects outside of the lab.

The effect modification issue implies that lab studies will misestimate the effects of decision-maker bias outside of the lab. But Cesario raises further concerns about the effect sizes claimed in the literature by comparing the path of interest (group membership → decision-maker bias → decision) to other paths (group membership → attributes of group members → decision). He implies that the latter explain much more variability in the final decision. We agree with Cesario that experimental studies on decision-maker bias following the standard design are not suitable to address this comparison.

This opens the door for the observational evidence relevant to the second path. Cesario states that "recent study suggests that the different rates of exposure to police through violent crime situations greatly – if not entirely – accounts for the overall *per capita* disparities in being fatally shot by the police" (sect. 4.1.2, para.3). which he uses as evidence that decision-maker bias is not to blame. Ross, Winterhalder, and McElreath (2021) show that this work incorrectly adjusts for crime rates. But, even if it had correctly adjusted for crime, if we formalize this claim about mediation; race → police exposure (e.g., in violent crime situations) → being fatally shot by the police, with only little or even no effects mediated through other pathways (captured in the remaining "direct effect," which would include decision-maker bias) remaining; we run into a problem (Fig. 1).

Exposure to police through violent crime situations will be affected by both race (including effects of earlier decision-maker bias) and other (potentially unobserved) factors (U). Conditioning on exposure induces collider bias, introducing spurious associations between race and U. For example, consider the possibility that Blacks are more likely to be involved with the police in general (e.g., Fryer, 2019) and that aggressiveness increases the chances to be confronted with the police (regardless of race). Without any actual group differences in aggressiveness, this means that – conditional on police exposure – Black individuals involved in such situations are less aggressive, which would decrease their chances of being fatally shot. Such induced confounding could hide decision-maker bias and has been discussed at great length (for summaries of the debate, see Hu, 2021; Lundberg, Johnson, & Stewart, 2021); it crops up for other topics as well (e.g., the gender wage gap, Hünermund, 2018). Outside of the lab, individuals are not randomly allocated to situations, which makes it challenging to identify decision-maker bias in observational data.

Perhaps the greatest benefit of an explicit causal inference framework is that it requires us to be more precise about the causal questions we are asking, thus enforcing conceptual consistency. Is Cesario trying to answer a forward causal question (the effect of decision-maker bias on outcomes) or a backward causal question (what causes group disparities in outcomes; Gelman & Imbens, 2013)? What counterfactuals are meant to be invoked? Counterfactuals about race ("What if this person were white instead of Black"), which have been criticized for being hard or impossible to define or otherwise inadequate (Kohler-Hausmann, 2018); or counterfactuals about racism ("What if there was no decision-maker bias"), as suggested by Krieger and Smith (2016)?

Clarifying these matters upfront may enable a more productive debate, as it ensures that we are not talking past each other. Causal inference frameworks do not magically guarantee value-free answers, but they force us to be precise about the questions we ask, and to be transparent about the assumptions that we are willing to make (e.g., Hu, 2021). This rigor is all the more important for politically charged topics where the stakes are high, and where it is all too easy to fall for clear-cut (counter) narratives.

Conflict of interest. None.

References

Fryer, R. G. (2019). An empirical analysis of racial differences in police use of force. *The Journal of Political Economy*, *127*(3), 1210–1261.

Gelman, A., & Imbens, G. (2013). Why ask why? Forward causal inference and reverse causal questions (No. w19614). *National Bureau of Economic Research*. https://doi.org/10.3386/w19614.

Grosz, M. P., Rohrer, J. M., & Thoemmes, F. (2020). The taboo against explicit causal inference in nonexperimental psychology. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, *15*(5), 1243–1255.

https://doi.org/10.1017/S0140525X21000017 Published online by Cambridge University Press

Hernán, M. A., & Robins, J. M. (2010). *Causal inference: What If*. CRC. https://grass.upc.edu/en/seminar/presentation-files/causal-inference/chapters-1-i-2/@@download/file/BookHernanRobinsCap1_2.pdf.
Hu, L. (2021). *Law, liberation, and causal inference*. LPE Project. https://lpeproject.org/blog/law-liberation-and-causal-inference/.
Hünermund, P. (2018). Nonlinear Mediation Analysis. https://p-hunermund.com/2018/01/17/nonlinear-mediation-analysis/.
Kohler-Hausmann, I. (2018). Eddie Murphy and the dangers of counterfactual causal thinking about detecting racial discrimination. Nw. UL Rev. https://heinonline.org/hol-cgi-bin/get_pdf.cgi?handle=hein.journals/illlr113&section=38.
Krieger, N., & Smith, G. D. (2016). The tale wagged by the DAG: Broadening the scope of causal inference and explanation for epidemiology. *International Journal of Epidemiology*, *45*(6), 1787–1808.
Little, R. J., & Rubin, D. B. (2000). Causal effects in clinical and epidemiological studies via potential outcomes: Concepts and analytical approaches. *Annual Review of Public Health*, *21*, 121–145.
Lundberg, I., Johnson, R., & Stewart, B. (2021). What is your estimand? Defining the target quantity connects statistical evidence to theory. *American Sociological Review*, *86*(3), 532–565. https://doi.org/10.31235/osf.io/ba67n.
Pearl, J., & Bareinboim, E. (2014). External validity: From do-calculus to transportability across populations. *Schweizerische Monatsschrift fur Zahnheilkunde = Revue Mensuelle Suisse D'odonto-Stomatologie/SSO*, *29*(4), 579–595.
Pearl, J., Glymour, M., & Jewell, N. P. (2016). *Causal inference in statistics: A primer*. John Wiley & Sons.
Rohrer, J. M. (2018). Thinking clearly about correlations and causation: Graphical causal models for observational data. *Advances in Methods and Practices in Psychological Science*, *1*(1), 27–42.
Ross, C. T., Winterhalder, B., & McElreath, R. (2021). Racial disparities in police use of deadly force against unarmed individuals persist after appropriately benchmarking shooting data on violent crime rates. *Social Psychological and Personality Science*, *12*(3), 323–332.

The importance of ecological validity, ultimate causation, and natural categories

Catherine A. Salmon and Jessica A. Hehman

Psychology Department, University of Redlands, Redlands, CA 92373, USA.
catherine_salmon@redlands.edu, jessica_hehman@redlands.edu

doi:10.1017/S0140525X21000741, e92

Abstract

The target article raises important questions about the applicability of experimental social psychology research on topics with policy implications. This commentary focuses on the importance of attending to a variety of factors to improve ecological validity as well as considering the ultimate factors shaping behavior and the role of natural categories in the stability of stereotypes and their influence.

We agree with the author's concerns about the flaws that are the focus of the target article. They are part of the ecological validity problem that plagues a variety of experimental approaches in the behavioral sciences (Salmon, 2020). While some psychologists have acknowledged concerns about generalizing from WEIRD (western, educated, industrialized, rich, and democratic) samples to the wider human population (Baumard & Sperber, 2010; Henrich, Heine, & Norenzayan, 2010), researchers have often failed to recognize the differences between the lab and the "real world" which challenges applicability to real world problems. The concern is whether there is a good fit between the task and the ecological problem it is approximating (Adolph, 2019; Salmon, 2020). While the target article focuses on decision-making, research examining effects of pornography on violence has also been plagued by experimental results that don't correspond to those of more ecologically valid studies (Diamond, Jozifkova, & Weiss, 2011; Ferguson & Hartley, 2020; Hatch et al., 2020)

When designing tasks to test hypotheses, we must be mindful of all contexts and factors the mechanism of interest is theoretically sensitive to and what aspects are likely to be more general across contexts. Consider deception detection research. The tasks participants face in detecting deception in the lab are quite different from those faced in everyday life, where they have a great deal more information about base rates of lying in their environment, specific individuals (reputation, non-verbal cues, and motivation), and ways to test if they are lying such as other sources of information (Levine, 2018). Lab studies are useful are in that they can provide information about the unique effect of an isolated factor on some particular outcome with "all else being equal." They are less helpful, however, for understanding how the system works when other relevant factors are included. Contrary to the strict experimental control employed in lab studies, in the real world, never is "all else equal." Individual and group differences exist, which likely influence behavior, attitudes, and cognition. Therefore, more multifactorial studies are needed to examine not only the effect of multiple factors at the same time, but also the potential combined effect of factors on some outcome.

Good science starts with thorough observations of the behavior of interest and then moves to hypothesis testing. The author highlights this in the police shooting case by pointing out the value of conducting task analysis of actual shootings first to get a more complete understanding of (a) all the relevant variables and (b) which may be most critical to include in experimental hypothesis testing. Assuming the variable you are interested in is the most relevant one, without comprehensive descriptive observational work, is likely to lead to erroneous conclusions for applications outside the lab. This is especially concerning when findings are used to inform public policy.

We disagree with the author that "distal causes of group differences are irrelevant" because they are "separable" from questions about specific outcomes. Rather than separate issues, they represent different levels of analysis. We would argue that another inherent problem of social psychological research has been ignoring ultimate factors (e.g., evolved psychological adaptations for solving recurring problems across our ancestral past) and focusing exclusively on the proximate mechanisms. Knowledge of *why* a particular behavior or outcome is occurring, informs our understanding of *what* is occurring as well as other factors that may be relevant.

The importance of understanding distal factors can be seen in the different patterns of stereotyping for different social groups. Although stereotype bias can be diminished by reducing ambiguity and providing more individuating information for some social groups, the reduction of ambiguity does not dramatically reduce stereotype bias against all social groups. Failure to acknowledge the ultimate origins of different types of stigma results in the error of assuming that all biased stereotyping processes occur under the same circumstances, in the same way, and thus may be mitigated in the same way. Consider differences between ageism and racism.

Consistent with age being a natural social category (i.e., one that would have existed across our evolutionary past), stereotypes about age tend to have cross-cultural similarities (Fiske, 2017). Across cultures, older adults are commonly described as being doddering but dear, incompetent but warm (Cuddy, Norton, &

Fiske, 2005; North & Fiske, 2015). The only cultures found to admire their older adults were Native Americans (Burkley, Durante, Fiske, Burkley, & Andrade, 2017) and African Americans (Fiske et al., 2009). Inconsistent with the popular belief that older adults are revered in Eastern cultures, older adults were found to be more derogated in Asian cultures than in Western cultures (North & Fiske, 2015). This is important in that it highlights a disconnect between attitudes toward older adults and cultural practices/expectations that aging parents are cared for by their adult children in Asian cultures. This disconnect suggests that stereotype beliefs about older adults do not necessarily map onto behavioral outcomes.

Stereotypes about social groups not considered to represent natural social categories (e.g., race), tend to be much more variable across cultures (Fiske, 2017). Different beliefs attached to varying ethnic groups depend on the specific nation and its history, including factors such as immigration history, income equality, political systems, and amount of conflict versus peace (Durante et al., 2017). Therefore, racial stereotypes appear to be determined by historical events (Fiske, 2017). The different patterns of racial stereotypes, for example, suggest race is being used as a proxy for something else (e.g., coalition membership). As pointed out by the author, racial stereotypes tend to be used in decision-making in the absence of other cues. A recent meta-analysis of "erasing race" studies in the United States confirmed a robust effect of reducing race-encoding by providing other cues to group membership such as team membership (Woodley of Menie et al., 2020). This suggests that racial biases may be activated under conditions of ambiguity, but not activated when other (perhaps more reliable) cues of coalitional membership are present. The same, however, does not apply to natural category biases. Those stereotypes appear to be consistently, unconsciously activated, even in unambiguous situations.

Financial support. The authors received no funding in support of this work.

Conflict of interest. The authors declare they have no conflict of interest.

References

Adolph, K. E. (2019). Ecological validity. In R. J. Sternberg (Ed.), *My biggest research mistake: Adventures and misadventures in psychological research* (pp. 187–190). Sage Publications.

Baumard, N., & Sperber, D. (2010). Weird people, yes, but also weird experiments. *Behavioral and Brain Sciences*, *33*(2–3), 84–85.

Burkley, E., Durante, F., Fiske, S. T., Burkley, M., & Andrade, A. (2017). Structure and content of native American stereotypic subgroups: Not just (ig)noble. *Cultural Diversity and Ethnic Minority Psychology*, *23*, 202–219.

Cuddy, A. J. C., Norton, M. I., & Fiske, S. T. (2005). This old stereotype: The pervasiveness and persistence of the elderly stereotype. *Journal of Social Issues*, *61*, 265–283.

Diamond, M., Jozifkova, E., & Weiss, P. (2011). Pornography and sex crimes in the Czech Republic. *Archives of Sexual Behavior*, *40*(5), 1037–1043. doi.org/10.1007/s10508-010-9696-y

Durante, F., Fiske, S. T., Gelfand, M., Crippa, F., Suttora, C, Stillwell, A., … Teymoori, A. (2017). Ambivalent stereotypes link to peace, conflict, and inequality across 38 nations. *Proceedings of the National Academy of Sciences, USA*, *114*, 669–674.

Ferguson, C. J., & Hartley, R. D. (2020). Pornography and sexual aggression: Can meta-analysis find a link? *Trauma, Violence, & Abuse*. doi.org/10.1177/1524838020942754

Fiske, S. T. (2017). Prejudices in cultural contexts: Shared stereotypes (gender, age) versus variable stereotypes (race, ethnicity, religion). *Perspectives on Psychological Science*, *12*(5), 791–799.

Fiske, S. T., Bergsieker, H. B., Russell, A. M., & Williams, L. (2009). Images of Black Americans: Then, "them," and now, "Obama!" *Du Bois Review: Social Science Research on Race*, *6*(1), 83–101.

Hatch, S. G., Esplin, C. R., Aaron, S. C., Dowdle, K. K., Fincham, F. D., Hatch, H. D., & Braithwaite, S. R. (2020). Does pornography consumption lead to intimate partner violence perpetration? Little evidence for temporal precedence. *The Canadian Journal of Human Sexuality*, *29*(3), 289–296.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*(2–3), 61–83.

Levine, T. R. (2018). Ecological validity and deception detection research design. *Communication Methods and Measures*, *12*(1), 45–54.

North, M. S., & Fiske, S. T. (2015). Modern attitudes toward older adults in the aging world: A cross-cultural meta-analysis. *Psychological Bulletin*, *141*, 993–1021.

Salmon, C. (2020). Multiple methodologies: Addressing ecological validity and conceptual replication. *Evolutionary Behavioral Sciences*, *14*(4), 373–378. doi.org/10.1037/ebs0000213

Woodley of Menie, M. A., Heeney, M. D., Peñaherrera-Aguirre, M., Sarraf, M. A., Banner, R., & Rindermann, H. (2020). A meta-analysis of the "erasing race" effect in the United States and some theoretical considerations. *Frontiers in Psychology*, *11*, 1635. doi.org/10.3389/fpsyg.2020.01635

How should we understand "bias" as a thick concept in recruitment discrimination studies?

Päivi Seppälä

Practical Philosophy, University of Helsinki, FI-00014 Helsinki, Finland.
paivi.a.seppala@helsinki.fi;
https://researchportal.helsinki.fi/en/persons/päivi-seppälä

doi:10.1017/S0140525X21000674, e93

Abstract

Cesario criticizes the experimental design of studies of bias by claiming that acting on stereotypes in the experimental situation might not be an "error" from a Bayesian perspective. However, social psychologists might have an ethical reason to label the observed decision-maker biases as "erroneous." Decision-making can be considered as "biased" and "erroneous," because it reflects illegal and morally condemnable discrimination.

Cesario's interpretation of experimental studies of bias in section 5 is correct as a standard Bayesian interpretation. Cesario rightly points out that experimental social psychologists fail to take properly into account that stereotypes may sometimes be accurate in everyday prediction. According to Cesario, acting on these stereotypes in the experimental situation might not be an "error" from a Bayesian perspective and experimental social psychologists should acknowledge this possibility.

However, social psychologists might have another, ethical, reason to label the observed decision-maker biases as "erroneous," and Cesario misses this reason when criticizing science, technology, engineering, and mathematics (STEM) hiring studies. In his criticism, Cesario does not acknowledge that researchers are not usually interested in merely explaining group disparities per se, but also aim to account for *discriminatory behavior* resulting in group disparities. Accounting for discriminatory behavior provides a rationale for the experimental design and the use of the word "erroneous" in the context of studying recruitment bias. Researchers labeling biased decision-making as "erroneous" do not merely claim that biased decision-making is wrong because it violates the norms of rational statistical inference in the context of the experiment. Instead, biased decision-making can also be labeled as an "error," because it results in illegal and morally condemnable discrimination. In this context, the words "discrimination" and "bias" are used as moralized concepts (Altman, 2020) or thick concepts (Williams, 1985) that simultaneously describe a phenomenon and express an evaluative stance toward it.

Let us consider the laboratory studies of STEM hiring in contrast to the goals of real-world hiring. In real-world hiring, the normative motivation for relying only on the relevant information provided by an applicant's resume is to prevent discrimination and to guarantee fair and equal treatment of applicants, which applicants also expect from recruiters (Gilliland, 1993). If one uses information on an applicant's membership in a salient social group as a decision-making criterion, this reasoning can be labeled as "biased, erroneous decision-making" because it violates the ethical norms of good recruitment practices. Following good recruitment practice, one judges a candidate based solely on the skills and merits of the applicant. This practice reflects the decision-making ideals of the recruiter, who wishes to closely adhere to the norms of anti-discrimination legislation (Koivunen, Ylöstalo, & Otonkorpi-Lehtoranta, 2015). These ideals are also widespread, because, for instance, gender discrimination in hiring is deemed illegal in 89% of countries (Heymann, Bose, Waisath, Raub, & McCormack, 2020).

In light of these norms, the experimental designs of STEM hiring are not mere displays of "methodological trickery," as Cesario suggests (sect. 5, para. 5). It is not trickery to create an experimental design where "the single relevant piece of information is the qualification of the applicant as revealed by the resume; being influenced by anything other than this information is treated as biased, erroneous decision-making" (sect. 5, para. 6). The design that Cesario describes reflects the real-world decision-making goals of recruiters and legislators. When a participant in a laboratory experiment uses irrelevant non-performance-related information on group membership to evaluate and to select candidates for an open position, the participant engages in "biased," "erroneous," and "discriminatory" decision-making that would count as "biased," "erroneous," and "discriminatory" decision-making also outside the lab.

Given that Cesario's goal is to suggest a new approach for experimental social psychology that begins with an analysis of actual decisions, Cesario should also walk the talk when criticizing STEM hiring research. The lesson is that participants in a laboratory study may be non-biased in the Bayesian sense, but at the same time their decision-making can be regarded as discriminatory and erroneous in the moral sense. First of all, it might be true that real-world recruiters (or college students enrolled in psychological experiments studying recruitment bias) might be Bayesian actors in the sense that they form their decision by using "information that may be probabilistically accurate in everyday life" (sect. 5, para. 7), as Cesario puts it. Second, it is also true that the experiments studying recruitment bias are designed in such a way that the label of "erroneous behavior" is attached to situations in which participants use information *within the experiment* that may actually lead to more accurate decisions *outside the experiment*. For instance, in some contexts, knowing that an applicant belongs to a certain salient social group might lead to somewhat accurate predictions about the applicant's future job performance or ability to commit to a job (Arrow, 1973; Phelps, 1972). Knowing about an applicant's childcare responsibilities might be a factor that has real-predictive value in some contexts. Nevertheless, the use of this information in a way that leads to disparate treatment of applicants in hiring decisions counts as statistical discrimination.

To conclude, it should be added that providing a deeper understanding of the motivation behind the experimental designs of STEM hiring does not show Cesario to be wrong in his main claim. One cannot naïvely assume that the social psychological experiments of categorical bias or audit studies provide causal explanations that would universally account for all real-world group disparities. What my comment puts forth is the possibility that the research methodologies of laboratory studies of recruitment and the interpretation of the results as "errors" may reflect the normative ethical values of the researchers and modern societies, because similar phenomena have occurred in other fields of science. Normative views on gender have been shown to influence how data are interpreted in anthropological studies on human evolution (Longino, 1990), and normative views of divorce have influenced the ways in which research questions and research designs are framed when studying the effects of divorce on well-being (Anderson, 2004).

It should also be noted that the purpose of my comment is entirely descriptive, and the goal is to correct and deepen Cesario's interpretation of studies of recruitment bias. I do not seek to defend the scientific soundness of the research methodologies and the ways of interpreting results by using the concepts of "bias" and "discrimination" as morally laden thick concepts. One can indeed question whether it is good scientific practice to allow values to influence science in this way.

Financial support. The study is funded by the University of Helsinki's 3-year research project "From cyborg origins of modern economics to its automated future. Towards a new philosophy of economics."

Conflict of interest. None.

References

Altman, A. (2020). Discrimination. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy: Vol. Winter 2020 edition*. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/entries/discrimination/.

Anderson, E. (2004). Uses of value judgments in science: A general argument, with lessons from a case study of feminist research on divorce. *Hypatia*, *19*(1), 1–24. https://doi.org/10.1111/j.1527-2001.2004.tb01266.x

Arrow, K. J. (1973). The theory of discrimination. In O. Ashenfelter & A. Rees (Eds.), *Discrimination in labor markets* (pp. 3–33). Princeton University Press.

Gilliland, S. W. (1993). The perceived fairness of selection systems: An organizational justice perspective. *The Academy of Management Review*, *18*(4), 694–734. https://doi.org/10.2307/258595

Heymann, J., Bose, B., Waisath, W., Raub, A., & McCormack, M. (2020). Legislative approaches to nondiscrimination at work: A comparative analysis across 13 groups in 193 countries. *Equality, Diversity and Inclusion: An International Journal*, *40*(3), 225–241. https://doi.org/10.1108/edi-10-2019-0259

Koivunen, T., Ylöstalo, H., & Otonkorpi-Lehtoranta, K. (2015). Informal practices of inequality in recruitment in Finland. *Nordic Journal of Working Life Studies*, *5*(3), 3–21.

Longino, H. E. (1990). *Science as social knowledge: Values and objectivity in scientific inquiry*. Princeton University Press.

Phelps, E. S. (1972). The statistical theory of racism and sexism. *The American Economic Review*, *62*(4), 659–661.

Williams, B. (1985). *Ethics and the limits of philosophy*. Fontana Press/Collins.

Cesario's framework for understanding group disparities is radically incomplete

Morgan Weaving and Cordelia Fine

School of Historical and Philosophical Studies, The University of Melbourne, Victoria 3010, Australia.
mweaving@student.unimelb.edu.au; cfine@unimelb.edu.au
https://findanexpert.unimelb.edu.au/profile/126041-cordelia-fine

doi:10.1017/S0140525X21000856, e94

Abstract

Cesario argues that experimental studies of bias tell us little about why group disparities exist. We argue that Cesario's alternative approach implicitly frames understanding of group disparities as a false binary between "bias" and "group differences." This, we suggest, will contribute little to our understanding of the complex dynamics that produce group disparities, and risks inappropriately rationalizing them.

Why do group-based inequalities exist? Cesario argues that the standard research paradigm in experimental studies of bias overstates the role of stereotypes (categorical bias) in decision-makers' perceptions and behavior. The harm in this, he suggests, is not merely the transmission of a "skewed … understanding of the human mind" (sect. 1, para. 3) to the wider culture, but the promotion of ineffective interventions – eliminating decision-maker bias – for addressing group disparities. Cesario's proposed alternative approach is: Detailed analysis of relevant decisions to ensure experimental tasks are valid representations of real-world processes; studying relevant "behavioral, personality, or other individual differences" (sect. 8, para. 3) between groups; and contrasting the effect size of categorical bias with other contributors to group disparities, particularly behavioral and personality group differences. We certainly agree that interventions aimed at eliminating decision-maker bias (e.g., blinding resumes) will not result in equal outcomes between groups. However, using gender disparities in labor market outcomes as an example, we disagree that Cesario's proposed approach will bring us closer to the goal of understanding or addressing group disparities.

Decades of scholarship have built an understanding of a gender system that (together with other interlocking hierarchical systems such as race and class) sustains inequalities of resources and authority via multiple, cumulative processes at the individual, interpersonal, institutional, and macro levels (Ridgeway & Correll, 2004). Within a gender system framework, then, the psychological processes identified by social psychologists simulate a single snapshot in time of just one of myriad interacting and dynamic mechanisms maintaining group disparities in status and resources. For this reason, we assume that there is broad consensus regarding Cesario's claim that simply eliminating decision-maker bias in any one particular context will not end group disparities in outcomes. Theorists of group inequalities have long recognized that formal equality on its own is inadequate to remedy the disadvantages of competing in a market in which the dominant group has already set the norms, practices, and standards (Young, 1990). *Contra* Cesario, we, therefore, doubt that many social psychologists believe that decision-maker bias alone can largely explain gender disparities.

Indeed, it's for this reason that research on the effects of stereotypes goes far beyond the "standard experimental approach" described by Cesario: from developmental psychology exploring the relations between toy exposure at home and gender stereotypical play (Boe & Woods, 2018); to psychobiological investigations of children's responsiveness to contrived gender cues and labels (Hines et al., 2016); to social psychological investigations of the effects of gendered stereotypes on career interest in science, technology, engineering, and mathematics (STEM) (Cheryan, Drury, & Vichayapai, 2013; Cheryan, Plaut, Davies, & Steele, 2009); to macro-level cross-national analysis showing that stronger gender stereotypes about mathematics among adolescents can explain cross-cultural variation in the gender gap in interest in a STEM career (Breda, Jouini, Napp, & Thebault, 2020).

Cesario acknowledges that there are many distal causes of group differences, potentially including social and structural ones, but regards these as, "irrelevant because these causes are separable from the question of whether group disparities are because of biased decision-making for specific outcomes. For example, the reasons why men and women differ in their interest in things versus people is a separate question from whether faculty search committees are biased against women in hiring for STEM positions" (sect. 1, para. 6). However, because of the complex ways in which these distal processes interact with and shape group differences (e.g., Stephens, Markus, & Fryberg, 2012), many social psychologists do not regard them as "irrelevant." Instead, they understand them as integral and interrelated parts of the system they are helping to unpack in their study of the effects of stereotypes.

Taking this broader view makes clear why Cesario's implicit framework, in which distal (and subsequent) effects are considered irrelevant, and whatever can't be explained by categorical bias is attributed to group differences in behavior, inadvertently encourages a false binary between "bias" and "group differences" as explanations of disparities. We agree that Cesario's suggestions will make for more accurate assessments of the contribution of decision-maker bias in a single decision-making context – and, in doing so, reduce the risks of allocating disproportionate resources to interventions likely to have modest or minimal effects. However, the research questions motivated by this implicit framework will not contribute much to understanding the complex dynamics that give rise to group disparities, and risk inappropriately rationalizing them.

Indeed, the latter point is illustrated by Cesario's stance that statistical discrimination (i.e., basing judgment of a member of a group in part on your "priors" about that group) is "a core tenet of good prediction" (sect. 5, para. 7). Thus, he is critical of the fact that "in studies of STEM hiring, the single relevant piece of information is the qualification of the applicant as revealed by the resume; being influenced by anything other than this information is treated as biased, erroneous decision-making" (sect. 5, para. 6). We are not exactly sure what other information (priors) Cesario thinks should influence recruiters. Is it the cross-culturally and ethnically variable sex difference in mathematical ability at the right-hand tail (Hyde, Lindberg, Linn, Ellis, & Williams, 2008; Penner, 2008)? Is it the gender gap in enjoyment of science among school-children that elsewhere is reversed (Stoet & Geary, 2018)? Is it men's lower contribution to unpaid labor (Hess, Ahmed, & Hayes, 2020) that provides them with more time to devote to advancement in a career in which long hours and the "zero-drag worker" are the norm (Williams, 2001; Williams & Smith, 2015)? Is it simply the fact that white men are the best represented demographic among scientists and engineers (National Science Board, 2018)? The use of any such priors in an employment context doesn't just risk a discrimination lawsuit. It also serves to rationalize the status quo, and to maintain and reproduce the reality of those priors.

A more detailed understanding of decision-maker judgments will not help our understanding of inequalities if it renders invisible other contributing factors and dynamics. What we need from social psychology is research investigating how psychological processes are shaped by, and contribute to, interpersonal, institutional, and macro-level factors that sustain group-based inequalities. The good news is such research is already flourishing.

https://doi.org/10.1017/S0140525X21000017 Published online by Cambridge University Press

Financial support. The authors received no financial support for the research, authorship, and/or publication of this article.

Conflict of interest. The authors declare no conflicts of interest with respect to the research, authorship, and/or publication of this article.

References

Boe, J. L., & Woods, R. J. (2018). Parents' influence on infants' gender-typed toy preferences. *Sex Roles 79*(5–6):358–373, https://doi.org/10.1007/s11199-017-0858-4.

Breda, T., Jouini, E., Napp, C., & Thebault, G. (2020). Gender stereotypes can explain the gender-equality paradox. *Proceedings of the National Academy of Sciences of the United States of America 117*(49):31063–31069, https://doi.org/10.1073/pnas.2008704117.

Cheryan, S., Drury, B. J., & Vichayapai, M. (2013). Enduring influence of stereotypical computer science role models on women's academic aspirations. *Psychology of Women Quarterly 37*(1):72–79, https://doi.org/10.1177/0361684312459328.

Cheryan, S., Plaut, V. C., Davies, P. G., & Steele, C. M. (2009). Ambient belonging: How stereotypical cues impact gender participation in computer science. *Journal of Personality and Social Psychology 97*(6):1045–1060, https://doi.org/10.1037/a0016239.

Hess, C., Ahmed, T., & Hayes, J. (2020). Providing Unpaid Household and Care Work in the United States: Uncovering Inequality (Briefing Paper No. IWPR #C487). Institute for Women's Policy Research. https://iwpr.org/wp-content/uploads/2020/01/IWPR-Providing-Unpaid-Household-and-Care-Work-inthe-United-States-Uncovering-Inequality.pdf.

Hines, M., Pasterski, V., Spencer, D., Neufeld, S., Patalay, P., Hindmarsh, P. C., & Acerini, C. L. (2016). Prenatal androgen exposure alters girls' responses to information indicating gender-appropriate behaviour. *Philosophical Transactions of the Royal Society B: Biological Sciences 371*, 20150125. http://dx.doi.org/10.1098/rstb.2015.0125.

Hyde, J. S., Lindberg, S. M., Linn, M. C., Ellis, A. B., & Williams, C. C. (2008). Gender similarities characterize math performance. *Science*, *321*(5888):494–495, http://dx.doi.org/10.1126/science.1160364.

National Science Board (2018) Science & Engineering Indicators. https://www.nsf.gov/statistics/2018/nsb20181/assets/nsb20181.pdf.

Penner, A. M. (2008). Gender differences in extreme mathematical achievement: An international perspective on biological and social factors. *American Journal of Sociology 114*(SUPPL. 1):138–170, https://doi.org/10.1086/589252.

Ridgeway, C. L., & Correll, S. J. (2004). Unpacking the gender system: A theoretical perspective on gender beliefs and social relations. *Gender and Society 18*(4):510–531, https://doi.org/10.1177/0891243204265269.

Stephens, N. M., Markus, H. R., & Fryberg, S. A. (2012). Social class disparities in health and education: Reducing inequality by applying a sociocultural self model of behavior. *Psychological Review*, *119*(4), 723–744. https://doi.org/10.1037/a0029028

Stoet, G., & Geary, D. C. (2018). The gender-equality paradox in science, technology, engineering, and mathematics education. *Psychological Science 29*(4):581–593, https://doi.org/10.1177/0956797617741719.

Williams, J. (2001) *Unbending gender: Why family and work conflict and what to do about it*. Oxford University Press.

Williams, J., & Smith, J. (2015) The Myth That Academic Science Isn't Biased Against Women. The Chronicle of Higher Education. https://www-chronicle-com.eu1.proxy.openathens.net/article/the-myth-that-academic-science-isnt-biased-against-women/?cid2=gen_login_refresh.

Young, I. M. (1990). Five faces of oppression. In I. M. Young (Ed.), *Justice and the politics of difference* (pp. 39–63). Princeton University Press.

Surely not all experimental studies of bias need abandoning?

Fiona A. White

School of Psychology, The University of Sydney, Sydney 2006, NSW, Australia.
fiona.white@sydney.edu.au
http://sydney.edu.au/science/people/fiona.white.php

doi:10.1017/S0140525X21000716, e95

Abstract

Cesario misrepresents experimental social psychology. The discipline encompasses significantly more than implicit bias research, including controlled decision making and real-world behavioral observations. Paradoxically, while critiquing popular implicit bias tasks, Cesario also describes task refinements that have significantly advanced their external validity and our contextual understanding of bias. Thus rather than abandonment, a call for "continued improvement" is a far more sensible proposition.

Race-, sex-, or religious-based bias, and their unsuccessful reduction, remain some of the most challenging areas for social scientists to conduct research. Such immutability may be a result of racial bias being analogous to a *social virus*, but unlike competing pandemic-based biological viruses, the structure and functioning of racial bias cannot be isolated under a microscope, and subsequently inoculated against via a society-wide vaccination roll-out. Racial bias is complex, it has both automatic and controlled processes, a multitude of moderating and mediating factors, and mutates according to the social context it infects. As a consequence of the ongoing challenges that racial bias presents to society and researchers who study it, a far more nuanced critique of this body of research than the one offered by Cesario's target article, is required.

I would like to begin my commentary with a strong defence of the strengths of experimental social psychology tradition, especially with its random assignment of participants to experimental and control conditions, the ability to make causal inferences compared to correlational research, and the potential to target these causes of bias and/or racism in prejudice-reduction interventions. Beyond acknowledging these empirical strengths, one concern lies with Cesario's somewhat naïve misrepresentation of experimental social psychology, where he reduces it to Greenwald, McGhee, & Schwartz's (1998) implicit association test (IAT), Correll, Park, Judd, & Wittenbrink's (2002) shooter bias task, and racial disparities in school disciplinary outcomes (using hypothetical vignettes). For the sake of accuracy, I would recommend that Cesario reframes his critique to the shooter bias and the IAT specifically, rather than using the broad phrase "experimental social psychology" throughout the target article. Without this necessary reframing, the article appears to ignore the many different and complex facets of experimental social psychology which also involves controlled decision making, real-world behavioral observations, applications, and longitudinal interventions. For example, one experimental social psychology area that has led the way in addressing the causes of racial bias is contact research, spurred on by Allport's (1954) original contact hypothesis formulations.

A wealth of contact research including experimental (White, Maunder, & Verrelli, 2020) and longitudinal (White & Abu-Rayya, 2012; White, Abu-Rayya, & Weitzel, 2014) research has shown that increased contact between members of different groups can reduce bias and prejudice and promote more positive intergroup relations. The prejudice-reducing effects of contact have been found to reduce anxiety consistently and robustly among both children and adults, and across many different types of contact settings and cultures (Tropp, White, Rucinski, & Tredoux, in press). For example, experimentally based E-contact between Catholics and Protestants in Northern Ireland, Muslim and Catholic students in Australia; people who identify as heterosexual and homosexual; people who identify as cisgender and transgender – where members of different groups engage in a structured, cooperative text-

based online discussion with one another – has been found to reduce anxiety and prejudice, and prepare individuals for direct outgroup contact (White et al., 2020). Clearly, experimental E-contact research *does* involve the presence of "real-world decision makers" from different religious, racial and sexual minority backgrounds, as well as "actual behavioral differences" being integrated into the E-contact paradigm, as evidenced by the structured synchronous-texting that occurs between each group member. This is one of many empirical examples that support my contention that *not all* experimental social psychology research of bias and prejudice should be tarred with the same brush of containing "fatal flaws" identified by Cesario. A more targeted critique is, therefore, warranted.

Another radical proposal by Cesario is that implicit bias experimental research should be "abandoned." Cesario's criticisms of poor ecological and predictive validity are not novel (see Blanton, Jaccard, Strauts, Mitchell, & Tetlock, 2015; Corneille & Hütter, 2020; Schimmack, 2021). According to the author, the shooter bias task is "fatally flawed" because it excludes necessary dispatch information, such as knowing the citizen's race beforehand. Consequently, when Johnson, Cesario, and Pleskac (2018) included these features in their laboratory experiment, the shooter bias effect disappeared. The refined Correll, Hudson, Guillermo, and Ma (2014) and Johnson et al.'s experiments clearly show that ecological improvements of the original paradigm impacted the levels of racial bias compared to what was initially reported. Cesario also claims that Blacks have higher violation rates than Whites, and that this ratio should be more accurately represented in the shooter bias task. Therefore, in addition to his methodological critique, the author presents a concerning undertone in his narrative that suggests that "with greater ecological validity White participants will report less racial bias… which is the truer picture," a narrative that seems counter to the evidence of the high number of fatal deaths of innocent Black men at the hands of White policemen. Overall, there appears to be an uneasy dissonance in the author's narrative, on the one hand he appears content with the shooter bias task's external validity when trained White police officers show no bias, but is critical of the task when racial bias is found among untrained participants.

Instead of supporting Cesario's extreme proposition to abandon the implicit bias experimental tradition, I propose that it should be "improved and refined" while racism continues to socially and emotionally infect intergroup relations globally. Cesario's "abandonment" position also ignores the significant contributions that the implicit bias research tradition, albeit flawed, has made to better understanding the predictors, mediators, and outcomes of racial bias. For example, showing bias exists, either directly or indirectly, is a necessary first-step in any researcher's attempts to effectively reduce it. In fact, there are several instances where even Cesario suggests possible refinements to the shooter bias task that can improve its external validity. Additional advances would be to examine these implicit bias tasks (measured in millisecond reaction time) alongside more refined controlled attitudinal and behavioral measures that focus on *intergroup* interactions, as per Allport's (1954) contact theory, rather than solely on *intragroup* processes. There remains academic merit in continuing to use *both* implicit and explicit approaches to examining bias within laboratory *and* fieldwork settings to better understand the causes, consequences, and the effective reduction of the *social virus* that is racial bias, and subsequently strengthen and promote the rigor of experimental social psychology.

Conflict of interest. None.

References

Allport, G. W. (1954). *The nature of prejudice*. Addison Wesley.

Blanton, H., Jaccard, J., Strauts, E., Mitchell, G., & Tetlock, P. E. (2015). Toward a meaningful metric of implicit prejudice. *Journal of Applied Psychology*, *100*, 1468–1481. https://doi.org/10.1037/a0038379.

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

Correll, J., Hudson, S. M., Guillermo, S., & Ma, D. S. (2014). The police officer's dilemma: A decade of research on racial bias in the decision to shoot. *Social and Personality Psychology Compass*, *8*, 201–213. http://dx.doi.org/10.1111/spc3.12099.

Correll, J., Park, B., Judd, C. M., & Wittenbrink, B. (2002). The police officer's dilemma: Using ethnicity to disambiguate potentially threatening individuals. *Journal of Personality and Social Psychology*, *83*(6), 1314–1329. https://doi.org/10.1037/0022-3514.83.6.1314.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology*, *74*(6), 1464–1480.

Johnson, D. J., Cesario, J., & Pleskac, T. J. (2018). How prior information and police experience impact decisions to shoot. *Journal of Personality and Social Psychology*, *115*, 601–623. https://doi.org/10.1037/pspa0000130.

Schimmack, U. (2021). The implicit association test: A method in search of a construct. *Perspectives on Psychological Science*, *16*(2), 396–414. https://doi.org/10.1177/1745691619863798.

Tropp, L. R., White, F. A., Rucinski, C., & Tredoux, C. (in press). Intergroup contact and prejudice reduction: Prospects and challenges in changing youth attitude. *Review of General Psychology*.

White, F. A., Abu-Rayya, H. M., & Weitzel, C. (2014). Achieving twelve-months of intergroup bias reduction: The dual identity-electronic contact (DIEC) experiment. *International Journal of Intercultural Relations*, *38*, 158–163. https://doi.org/10.1016/j.ijintrel.2013.08.002.

White, F. A., & Abu-Rayya, H. M. (2012). A dual identity-electronic contact (DIEC) experiment promoting short- and long-term intergroup harmony. *Journal of Experimental Social Psychology*, *48*, 597–608. https://doi.org/10.1016/j.jesp.2012.01.007.

White, F. A., Maunder, R., & Verrelli, S. (2020). Text-based E-contact: Harnessing cooperative internet interactions to bridge the social and psychological divide. *European Review of Social Psychology*, *31*(1), 76–119. https://doi.org/10.1080/10463283.2020.1753459.

Author's Response

Reply to the commentaries: A radical revision of experimental social psychology is still needed

Joseph Cesario

Department of Psychology, Michigan State University, East Lansing, MI 48824, USA.
cesario@msu.edu
www.cesariolab.com

doi:10.1017/S0140525X21002405, e96

Abstract

Are the landscapes of real-world decisions adequately represented in our laboratory tasks? Are the goals and expertise of experimental participants the same as real-world decision-makers? Are we neglecting crucial forces that lead to group

outcomes? Are the contingencies necessary for producing experimental demonstrations of bias present in the real world? In the target article, I argued that the answers to these questions are needed to understand whether and how laboratory research can inform real-world group disparities. Most of the commentaries defending experimental social psychology neglected to directly address these main arguments. The commentaries defending implicit bias only revealed the inadequacy of this concept for explaining group disparities. The major conclusions from the target article remain intact, suggesting that experimental social psychology must undergo major changes to contribute to our understanding of group disparities.

R1. Introduction

I group the commentaries into four categories: (R2) commentaries that build productively on the target article; (R3) critical commentaries that address the main arguments in the target article; (R4) commentaries that misunderstand the target article; and (R5) commentaries that attempt to salvage the concept of implicit bias.

R2. Productive discussion and directions

R2.1. Big picture contributions

The standout commentary was by **Rohrer, Schmukle, and McElreath** (**Rohrer et al.**), who identified the main weakness of the target article: "lists of problems are not solutions." Rohrer and her colleagues raise questions regarding causal inference that must be addressed in order to study bias and disparities (see also Cesario, 2021 and Ross et al., 2021) and correctly note the wide applicability of their approach. Indeed, although they focus on past research that has questioned the existence of racial bias, the same problems apply equally to research that purports to show evidence of categorical bias.

It is useful to explore the overlap between the causal inference approach and traditional experimental approaches. Scholars have noted that causal models are built from existing scientific knowledge, and one source of such knowledge is experimental data (e.g., McElreath, 2021). In building a causal model one must make starting-point decisions about which variables to include and how they are related, and these decisions are informed by the scientist's assessment of existing data. If such data come from flawed experiments, however, then the causal models themselves may be mis-specified. Causal inference will not be the savior we need if such models are based on data from the kinds of social psychology experiments criticized in the target article.

I gently push back on **Rohrer et al.**'s reduction of the target article to one of mere "effect modification." Similarly, **Mora, Klein, Leys, and Smeding** (**Mora et al.**) suggest that the missing flaws could be reconceptualized as moderating variables. These authors are correct that many of the problems I raised can be subsumed under the idea that some feature of the decision landscape moderates an effect. But it would be a mistake to lose the broader critique about how the research strategy of experimental social psychologists leads them to fundamentally misunderstand the topic under study and produce a skewed view of human nature.

Białek and Grossmann raise the important distinction between judgment and decision-making. They note that judgments "are often decontextualized, are made on the fly, and bear little consequences to the agent." But to what degree are important group disparities because of such judgments? Choosing to eat pulled pork is one thing; it is something else entirely to hire a programmer or reject a mortgage application. Is the lending agent at a bank making a decontextualized decision on the fly, with no consequences to herself? Consistent with the target article, Białek and Grossmann's distinction requires starting with a task analysis of the decision so that we better understand its nature, including the informational inputs at play.

Mitchell and Tetlock argue that failure to attend to construct and external validity has undermined the quality of psychological science and public trust in our findings. They note "if the goal of social psychology is to create an ideology…then it appears to have succeeded." Although I mostly stayed away from ideological speculations in the target article, I agree with these authors and suggest that increased political and ideological diversity will improve many of the problems of experimental social psychology. I turn to this topic next.

R2.2. Importance of heterodoxy and diversity in science

I interpreted a number of the commentaries as supporting the argument for increased political and ideological diversity in psychological science, both for researchers and participants (consistent with **Matsick, Oswald, and Kruk**'s and **Hodson**'s commentaries).

Rini suggests that I made "overly skeptical demands" of social psychology. Although she is appropriately cautious about the nascent state of our knowledge, other researchers have been nowhere near this careful. For example, proposing that U.S. legal doctrine be changed on the basis of implicit bias work, before even a coherent and empirically supported definition of the term was established, in no way resembles Rini's more reasonable position. I contend that my demands are not out of proportion given the overt political activism among social psychologists.

This problem of premature use of experimental findings supports the call for greater political diversity in academia. Strong political beliefs can lead to an "ends justify the means" mentality in which questionable research is given a pass because it meets one's desired political goals. Political and ideological diversity help mitigate the problems of motivated reasoning and selective calls for rigor.

I agree with **Brownstein, Kelly, and Madva** (**Brownstein et al.**) that I neglected the social and situational forces acting on researchers as they plan and carry out their research programs. Concerns about the social norms that allow researchers to conduct and publish flawed research support the importance of ideological diversity. Such diversity can help prevent groupthink and acceptance of the status quo by providing constant challenges to the methodological assumptions that underlie research programs.

Hodson argues that "we should be mindful" of how research findings will be used, which is a step away from subordinating the scientific process to political concerns. We should be in the business of truth-uncovering, not worrying about whether White supremacists or Black Lives Matter activists will better like our findings. For example, it is critical to understand how people's own behavior plays a role in police shootings if we are to enact policies that reduce racial disparities; presupposing that they do not ("nonsensical," according to **Blake**) and ignoring a causal analysis of shootings will not serve this goal. It will only serve political ends, and there is nothing in political success that necessarily translates to a better life for people other than politicians and activists.

https://doi.org/10.1017/S0140525X21000017 Published online by Cambridge University Press

Hodson also bemoans the shift toward "right-leaning priorities," as if the concerns of roughly half the U.S. population should not be of interest to the discipline of social psychology. I do think he raises important questions about how to define prejudice and who controls that definition. But such an answer (effectively, only those on the political left) is evidence that we need more political diversity in science. Moreover, being surrounded only by political allies makes it easy to believe that our good intentions will lead to good outcomes in our quest to shape the world according to our vision (Sowell, 1983, 1999, 2015; Williams, 1982). But without diverse perspectives to offer constant challenge, it is easy to miss the costs and unintended negative consequences inherent in any political solution (e.g., Devi and Fryer, 2020). Political diversity could also mitigate the unintended costs raised by **Arkes**, insofar as people from different political positions are more or less sensitive to different types of costs.

Seppälä argues that although the use of categorical information in experiments can be accurate from a Bayesian standpoint, researchers are correct in calling such decisions errors because the use of categorical information is "morally condemnable." She states (emphasis added):

> If one uses information on an applicant's membership in a salient social group as a decision-making criterion, this reasoning can be labeled 'biased, erroneous decision-making'…one judges a candidate based *solely* on the skills and merits of the applicant.

By **Seppälä**'s logic, affirmative-action selection and "diversity" hiring decisions are biased, erroneous, and morally condemnable because they involve judging a candidate on something other than "skills and merit." Whether such selection is viewed as morally condemnable or morally righteous clearly depends on one's political and ideological standpoint, and thus her commentary ultimately argues in favor of increasing political and ideological diversity.

Jasperse, Stillerman, and Amodio (**Jasperse et al.**) accuse me of "modern racism" for believing that one source of disparate outcomes is the behavioral differences across groups and for leaving open the possibility that such differences are inherent to people. First, I said that behavioral differences were one of the many factors producing disparate outcomes. Second, there is little definitive evidence about the ultimate causes of group differences and therefore the most defensible position is one of agnosticism rather than certainty. Political and ideological diversity is needed in academia to allow such questions to continue to be asked, because people from different positions will have different starting priors and will differ in the questions they consider to be "acceptable" in the first place.

Qu-Lee and Balcetis provide a compelling account of how differences in visual processing may represent an important source of bias. Relatedly, **Oyserman and Jeon** provide a framework for thinking about cultural variation and bias. What is unclear is whether research programs based on these frameworks would withstand the flaws identified in the target article. For example, understanding when bias may be more or less likely is a different goal than explaining group disparities, and applying the cultural fluency or the visual processing framework seems better suited to the former rather than the latter.

The commentary by **Ledgerwood, Pickett, Navarro, Remedios, and Lewis** (**Ledgerwood et al.**) makes a great case for increasing political and ideological diversity in psychological science. (I address their specific arguments below.) To the extent that researchers have similar ideological and political beliefs, even if they are demographically diverse, the benefits of team science are less likely to be realized.

R2.3. Building on the target article

A number of commentaries extended the arguments from the target article in interesting ways; space prevents me from going into detail on these. **Arkes** provides examples of important "unintended costs" of using experimental social psychology to understand group disparities. **Biggs** argues that the failure of experimental studies to include anticipated consequences, especially in life-or-death situations, makes these studies even less informative than described in the target article. **Blake** provides additional points which could serve as a counter-argument to other commentaries ("the idea of ignoring an individual's antecedent behaviors proximal to a police shooting is, bluntly, nonsensical").

Burt and Boutwell extend the target article in a useful way to other methods of investigation, including self-report of discrimination experiences. **Salmon and Hehman** build a convincing case for the importance of incorporating ultimate causes in understanding stereotyping processes and outcomes. More proximally, **Rennels and Insouvanh**, coming from a developmental view, connect the target article to a similar critique by Dahl (2017) and show how naturalistic observations can address the flaws identified in the target article.

R3. Commentaries critical of the target article

R3.1. Commentaries suggesting flaws or errors in my analysis

Freeman, Johnson, and Stroessner (**Freeman et al.**) make the important point that accuracy and bias can co-exist in decision-making. To the extent that the target article "implies a zero-sum tradeoff between accuracy and bias," my writing should have been clearer. Relatedly, **Jasperse et al.** argue that the missing features identified in the target article may have an amplifying rather than attenuating role in the expression of racial bias.

In highlighting how bias and accuracy in decision-making can have a complex relationship, both **Freeman et al.** and **Jasperse et al.** support the major conclusion of the target article: The jump from experimental demonstrations of decision-maker bias to explaining group disparities is misguided. All told, we are left with a situation in which experimental demonstrations of bias can be enhanced, eliminated, reversed, or unaffected by the missing forces identified in the target article. It is not clear how studies of experimental bias contribute to group disparities without doing the kind of work I advocated in the target article. That the relationship between bias and accuracy, or between bias and context, might be *more* complex than the one I focused on in the target article hardly saves experimental social psychology.

Tellingly, although **Jasperse et al.** claim that missing context can amplify racial bias observed in experimental studies, they provide no empirical evidence supporting this claim. These authors cite "systemic" policing factors as ways in which experimental biases may be amplified. First, I was clear that the target article concerned decision-maker bias and that factors "earlier" in the chain can also be a source of bias; Jasperse et al. simply cite one of these earlier factors. Second, far from supporting the claim that systemic factors amplify the biases observed in the lab, the work cited by these authors supports the target article

https://doi.org/10.1017/S0140525X21000017 Published online by Cambridge University Press

by demonstrating how we might *incorrectly* attribute outcome disparities to decision-maker bias when they instead stem from factors other than individual-level biases. (For a similar error, see the section on **Fuentes, Ralph, and Roberts** [**Fuentes et al.**] below.) Even when there is no bias in officers' decisions to shoot, such factors can produce racial disparities, which was a point I made in the target article and elsewhere (Cesario, 2020; Cesario, Johnson, & Terrill, 2019).

Duell and Landa argue that there is still utility in knowing how actors respond to hypothetical situations, even if such decision situations do not occur in the world. On the one hand, they provide a convincing argument for the importance of knowing how people would respond to some theoretical situation and that experiments can reveal anticipatory discrimination. On the other hand, if one thinks that audit studies may reveal something about, say, underinvestment in potential Black employees, why not study that directly?

Essien, Stelter, Rohmann, and Degner (**Essien et al.**) correctly note that the target article was not concerned with intergroup prejudice, and they suggest that experimental demonstrations of intergroup prejudice might shed light on group disparities. Yet while intergroup prejudice is certainly widespread, the degree to which prejudice can explain group *disparities* is far from obvious. Indeed, cross-cultural and historical study provide ample evidence that liking is not required for achievement (see, e.g., Sowell, 2019). From the economic achievement of the disliked "Overseas Chinese" throughout SE Asia, to the success of Jewish people facing strong dislike and discrimination across the globe and across history, to the economic outperformance of U.S. Whites by Japanese-, Chinese-, Indian-, and Caribbean-Americans, to the disconnect between levels of prejudice and the variation in achievement among the U.S. White ethnic groups, the link between prejudice and disparities is far from straightforward.

Hodson argues that the consistency between experimental results and real-world analyses renders the target article impotent. First, a stopped clock is still right twice a day. Second, he overstates the quality of the evidence supporting his claim. For example, he cites two papers on audit studies in the labor market, but in the target article I had already addressed the problems with using such studies to understand group disparities. For police shootings, which was the focus of the target article, he presents no evidence of racial bias by police officers in the decision to shoot, instead citing other, related work.

R3.2. Commentaries advocating to "stay the course"

White; **Duell and Landa**; **Rini**; **Seppälä**; and **Okonofua** all make some version of the argument that a productive way forward is to improve our current experimental and inferential cycle rather than radically change experimental social psychology. I disagree. If we continue to begin our investigation with the assumptions in the heads of social psychologists rather than with a task analysis of the decision itself, we may keep missing critical parts of the decision landscape and misunderstand the outcomes of interest.

Okonofua claims that experimental work on school disciplinary disparities has already addressed the critical flaws outlined in the target article. Okonofua overstates the quality of the evidence from experimental studies of school disciplinary disparities and in some cases misses the point of the target article entirely. For instance, he claims that my missing information critique "lacks factual merit" because his stimuli are representative. But the argument was that the scenarios are impoverished, not that they are unrepresentative; that is, experimental decision-makers lack information about specific students that is available to real teachers. It is undeniable that a one-paragraph description of a stranger is impoverished relative to the knowledge accumulated day after day in a school classroom.

Okonofua further points to his work on "two-strikes" as a means of defending experimental studies from the "missing information" flaw. He argues that adding information about a child's history of misbehavior increased racial bias; hence my missing information flaw is not only inapplicable but exactly wrong. In that work by Okonofua and Eberhardt (2015), there was no racial bias when it was a child's first infraction, but teachers treated Black students more harshly when it was the child's second infraction (i.e., bias increased when more information was added to the decision).

A closer examination of the quality of this work suggests that **Okonofua** is too optimistic in his defense of experimental psychology. First, this effect failed to replicate in a preregistered replication (Jarvis & Okonofua, 2020). Second, underlying the descriptive, verbal claim of "teacher bias" are important inconsistencies across the three relevant studies on the topic. Across eight different dependent variables in Okonofua and Eberhardt (2015) and Jarvis and Okonofua (2020), some dependent variables show race effects and others do not – but it is a *different set* of supportive and unsupportive variables in each study. The variability present in the actually obtained, specific effects (rather than just general claims of evidence of "bias"), paired with the small sample sizes and p-values close to 0.05, do not lead to confidence that the existing experimental studies provide strong evidence on the question of racial disparities in disciplinary outcomes.

On the question of behavioral differences across groups, **Okonofua** rightly points out that he has done work exploring this issue (which I discussed favorably in the target article). Yet he also points out that "student misbehavior cannot fully account for racial disparities in discipline." This is an important point to explore for several reasons. First, the argument of the target article was not that behavioral differences could account for 100% of racial disparities in disciplinary outcomes, just that such differences were not appropriately considered by experimental studies on the topic. This leads to little-discussed question of "how much": How much of the disparity do social psychologists hope to explain? If, for example, teacher bias accounts for 1% of the racial disparity in disciplinary outcomes once all other factors are considered, is this worth the costs of studying the topic and of implementing interventions designed to reduce teacher bias? Costs include both taxpayer-funded grants into research and interventions but also the unintended consequences associated with any intervention (e.g., Anderson, Ritter, & Zamarro, 2017), including public misunderstanding of the size of such effects and the role bias plays in producing group disparities.

Second, **Okonofua** notes that education researchers have also come to the conclusion that racial disparities "cannot be solely attributed" to differences in misbehavior, citing a review by Welsh and Little (2018). While technically true, this is hardly a strong conclusion from the research cited in that review. Welsh and Little do claim that "misbehavior…does not fully explain the rates of or disparities in exclusionary discipline outcomes" (p. 757) and cite Skiba et al. (2014) as evidence for this claim. Yet Skiba et al. is a database of only those students who have obtained the outcome (i.e., who have been suspended or expelled); as has been pointed out in other contexts (e.g., Knox and

https://doi.org/10.1017/S0140525X21000017 Published online by Cambridge University Press

Mummolo, 2020), this prevents clear inferences about the causal forces that do and do not produce the outcome.

Even so, Skiba et al. showed at best small race effects comparing in-school versus out-of-school suspensions, with "Black students being more likely to receive OSS (OR = 1.248) than White students." Moreover, this small odds ratio of 1.2 was reduced to non-significance once school characteristics were added, which "suggests that racial disparities in the use of out-of-school suspension may be explainable by a range of school-level variables." This and other complications led Welsh and Little to note that "there is little empirical evidence to substantiate the notion that discriminatory behavior by teachers and school leaders is a significant driver of discipline disparities" (p. 758).

R3.3. Commentaries highlighting the role of "systemic" or "structural" factors

A number of commentaries argued that "structural" or "systemic" bias was not appropriately treated in the target article. **Fuentes et al.** claim that I created "an artificial line" between decision-maker bias and systemic bias. Instead, they intentionally blur an otherwise-clear distinction, which allows them to suggest that my *true* rationale for writing the target article was to undermine the "convergence" across the social sciences on the importance of systemic factors. (In fact the only "convergence" is mere ideological homogeneity and not empirical consistency.)

Fuentes et al. raise the possibility that structural features can bias individual decision-making. I did not deny this possibility in the target article; but this still places decision-maker bias as a key factor in producing group disparities. The question then remains as to whether experiments can reveal these biases in meaningful ways, that is, we revert back to the main arguments of the target article.

That said, let us evaluate the quality of the evidence provided by **Fuentes et al.** in support of their argument that "structural features can bias individual decision-making." The target article is straightforward: One missing element in the experimental approach to racial disparities in fatal police shootings is the difference in violent crime rates across racial groups. There is evidence that exposure to the police via violent crime contributes meaningfully to such disparities. Failure to consider this leads to potentially unwarranted claims about officer racial bias in the decision to shoot. What evidence do Fuentes et al. provide to undermine this argument? They suggest that I fail to "take into account that the reason Black individuals encounter police at higher rates is largely because police departments target segregated Black neighborhoods for greater surveillance and intervention. Police violence is *structured* to impact Black individuals more."

Before tackling the specific citations provided for these claims, it is important to note that **Fuentes et al.** omit a critical element necessary to understand differential deployment: the *different crime rates* found in different neighborhoods and among different racial groups (see, e.g., Latzer, 2018, 2020). This is not to say that deployment is perfectly calibrated to crime rates. But Fuentes et al. obscure the nature of differential policing with their framing of a *structural* bias as "largely" the reason without including the very real crime rate differences that correspond to differential policing. The mere fact of differential deployment is taken as a "structural bias" intentionally designed to disproportionately impact Black Americans. This structural argument is made with no data or discussion about the *degree* of miscalibration, nor a discussion about what type of deployment patterns we would expect in the absence of bias, nor with any citation supporting the claim that differential deployment is "largely" because of intentional targeting in a manner divorced from crime rate differences.

As evidence that I have misunderstood the nature of fatal police shootings and the role of violent crime differences across racial groups, **Fuentes et al.** cite five sources. First, they reference the FBI Uniform Crime Report data. They do not provide any detail about how exactly the FBI data undermine my argument; nor can they, because I used FBI UCR data in making the original argument (as well as other crime rate sources; see Cesario et al., 2019). They are wrong in stating that the UCR data do not show racial differences in crime rates. Second, they cite Dunham & Petersen (2017), which is a policy recommendation paper that provides no evidence that violent crime rates do not contribute to disparities in fatal shootings. Third, they cite Hehman et al. (2018); I discuss this paper in section R5. Fourth, they cite Swencionis & Goff (2017), which is a review paper that provides no evidence that violent crime rates do not contribute to disparities in fatal shootings. Finally, they cite Gordon (2020), which has no supportive data on the claim that deployment of police to different neighborhoods does not reflect crime rate differences across those neighborhoods.

Hence, none of the cited evidence by **Fuentes et al.** is convincing.

Even more, **Fuentes et al.**'s own argument (however lacking in empirical support) undermines their defense of implicit bias. If police are structurally targeting Black neighborhoods, then implicit bias on the part of individual officers making deadly force decisions is not needed to explain racial disparities in that outcome.

In the end, it is worth remembering a complementary lesson to be taken from the commentary by **Arkes**. While the target article argued that sometimes decision-maker bias is not what it seems, sometimes "systemic" bias is not what it seems either.

R4. Misunderstandings/misattributions

R4.1. Commentaries which misunderstood the target article or attributed incorrect claims to it

Freeman et al. misattribute claims to the target article. For instance, they state that I implied "accuracy in decision-making obviates bias or the need to study it," "investigating bias when people are generally accurate is unnecessary," and that "target-driven differences between groups…invalidates decision-makers' bias or the need to study it." I never made any of these claims.

More fundamentally, **Freeman et al.** suggest that I "misrepresent" research because "few studies explicitly explore the link between implicit bias and real-world group disparities. Instead, most bias research aims to document group-based distinctions in individuals' decisions." Yet I was clear that experimental studies of bias "can and do tell us about the functions and processes of storing group-based information" and that my concern was with the application of such research to understanding group disparities. As if to perfectly prove my point, Freeman et al. make exactly this application in the very next sentence by claiming that such studies are important "because demonstrating such a bias illuminates one factor contributing to gender-based differences in science, technology, engineering, and mathematics (STEM) representation." Yet this is precisely the problem discussed in

https://doi.org/10.1017/S0140525X21000017 Published online by Cambridge University Press

the target article: that the flaws of experiments prevent such an application. In other words, they make exactly the error I accuse social psychologists of making and which they claim is a strawman.

Jasperse et al. accuse me of "misrepresenting" the findings of Correll et al. (2011), yet it is not obvious how I misrepresented this finding. Placing targets in dangerous backgrounds eliminated racial bias in the decision to shoot, for whatever reason (whether because of direct effects of criminality of the environments or because of the "racially coded" nature of those environments). Similarly, Jasperse et al. claim that the Moss-Racusin et al. (2012) study "uses none of the methods he critiques." This is baffling given that this is a field audit study in the tradition of the labor market studies discussed in the target article.

To clear up two misunderstandings revealed by **Weaving and Fine**'s commentary. First, I did not mean to suggest that decision-maker bias and group behavioral differences are locked in a zero-sum relationship. The argument was that experimental studies of decision-maker bias cannot tell us about disparities (in part) because they fail to incorporate group differences. Second, it was not my intention to suggest that distal factors are irrelevant to *understanding* group disparities. In referring to them as "irrelevant," I was stating they were irrelevant to the specific argument in the target article about whether experimental studies of decision-maker bias inform group disparities.

Ledgerwood et al. suggest that the target article itself contained three flaws which undermine my argument. None of these flaws hold up and in some cases the authors misunderstand the argument or are simply incorrect about their claims. Their first flaw is *biased search*; here, the authors claim that my expectations led me to a biased search of the existing literature and that I missed crucial information that was inconsistent with my thesis. For example, they argue that I missed the possibility that changes to the parameters of an experiment can amplify rather than eliminate bias. At no point did I suggest that discrimination or bias was absent in the real world or that the magnitude of such discrimination might be greater than that found in the lab; the target article questioned whether experimental findings could be used to understand group disparities. Thus, these authors miss the main point and instead attack a different claim, one not made anywhere in the target article. As if to exactly prove my point, they cite studies that have nothing to do with experimental social psychology and in no way demonstrate that adding missing information amplifies the effects observed in our experiments.

As a second example, they argue that additional information can sometimes sustain and justify bias. They reference Darley and Gross (1983), in which participants ($N = 14$ per cell) gave ratings in line with their initial positive or negative expectation only when presented with ambiguous performance information. Nothing in the target article suggested that people cannot engage in motivated reasoning. The issue concerns the degree to which our experimental situations match the *real-world decision scenarios* to which experimental findings are applied. Let us look in detail at the materials from Darley and Gross. The key performance information is that the target:

> answered both easy and difficult questions correctly as well as incorrectly. She appeared to be fairly verbal, motivated, and attentive on some portions of the tape and unresponsive and distracted on other portions of the tape. The tester provided little feedback about Hannah's performance.

Perhaps in a world governed by the "power of the situation," people are equally likely to answer difficult and easy questions correctly or vacillate between being attentive and unresponsive from one moment to the next. Perhaps, teachers can gain no useful feedback about their students' performance. For these situations, I concede that **Ledgerwood et al.** are correct – as I already did in the target article where I emphasized the importance of ambiguous, non-diagnostic information for categorical bias to dominate.

A concrete example may help illustrate the irrelevance of experimental findings such as Darley and Gross for explaining real-world disparities. In 2017, there were 13 high schools in the city of Baltimore (student population: over 85% Black, under 5% White) with *zero* students who tested proficient at grade level in math. In Baltimore's Augusta Fells High School, 50% of students in 2020 had a grade point average (GPA) of *0.13 or lower*. This is not the kind of ambiguous, non-diagnostic performance that **Ledgerwood et al**. would suppose exists in pointing to Darley and Gross as an example of how real-world disparities can be elucidated by experimental social psychology. To suppose that the decision situation of Darley and Gross somehow matches the conditions found in examples like these is not an idea to be taken seriously and there is little reason to predict that policies based on findings such as Darley and Gross will do much to reduce racial disparities in this outcome.

Thus, while **Ledgerwood et al.** claim that my biases caused me to miss some crucial data, they do not provide any evidence supporting that claim.

The second flaw described by **Ledgerwood et al.** is the *Beginner's Bubble Flaw*, which they use to suggest that I overestimated how well I understood the topics discussed in the target article. Specifically, the authors take issue with my discussion of the Bayesian framework to understand stereotype use. Here, the authors attribute misleading and false statements to the target article, arguing against points I never made. They state that I claimed "using demographic information (e.g., race) to fill in the blanks when full information is unavailable is rational in a Bayesian sense and therefore unbiased." Here, is the relevant passage in the target article:

> Such a demand on the part of social psychologists in fact violates a core tenet of good prediction, which is the use of priors in updating posterior prediction. Bayes' rule would require participants in social psychology experiments to include the target's categorical information in their judgments (though of course the effect of categorical information should depend on the strength of the data, as it does). (sect. 5, para. 7)

My argument was that *in the context of social psychology experiments*, researchers demand decisions that would violate Bayes' rule; that is, they require participants to use no prior information and to treat all targets in an identical way when no diagnostic information is provided. Such a demand simply does not make sense from a Bayesian standpoint.

I did not say that using priors is "unbiased." I did not say that using priors is "rational or justifiable." I did not say that a prior is "the same thing as a base rate…the same thing as truth." I did not say that "just because a belief can *sometimes* lead to correct decisions…it is accurate or optimal to use that belief for all decisions." In contrast to **Ledgerwood et al.**, both **Arkes** and **Seppälä** (in their solo-authored commentaries) correctly understood my discussion of this issue.

https://doi.org/10.1017/S0140525X21000017 Published online by Cambridge University Press

With their final flaw, *Old Wine in New Bottles*, **Ledgerwood et al.** suggest that I did not "recognize prior work" that has connected the lab and the real world. They list a number of citations to suggest that the target article is covering old ground. Let us look at each of these to see whether the criticism holds. Aronson and Carlsmith (1968) are unrelated to understanding group disparities, but regardless, I did acknowledge related research on experimental realism and mundane realism. Bauer, Damschroder, Hagedorn, Smith, and Kilbourne (2015) concerns best practices for implementing evidence-based practices into applied settings; the target article instead concerns whether we should be doing that in the first place. IJzerman et al. (2020) was published while the target article was under review, and so is contemporaneous with the target article and not "prior work." Premachandra and Lewis (2021) are about reporting practices for intervention studies, which is unrelated to the argument from the target article; it also appeared after the target article was accepted. Lewin (1946) is an excellent article on social problems from a scientific approach and could have strengthened section 8 of the target article. So in sum, according to Ledgerwood et al., I missed one paper related to a single section of the target article.

Finally, **Brownstein et al.; Jetten, Selvanathan, Crimston, Bentley, and Haslam (Jetten et al.)**; **Kurdi and Dunham**, and **White** all offer some version of "yes, but experiments can do lots of things." White notes that experimental social psychology "encompasses significantly more than implicit bias research." Kurdi and Dunham (incorrectly) claim that I stated all research seeks to explain group disparities, and that my "misleadingly narrow" claim misses out on my research areas such as "memory research" and "phonological awareness research." Brownstein et al. highlight the "experimental and theoretical" improvements that have been accomplished by social psychologists and argue that the blank slate worldism of experimental psychology "is entirely appropriate *for the epistemic aim* of determining that bias exists." Finally, Jetten et al. claim that the target article is flawed because it "overlooks the importance of theory testing."

None of these are convincing counter-arguments to the target article. Of course experiments can do many things and social psychology covers a range of topics. The target article was about applying experimental social psychology to group disparities. Pointing out that experiments can also be conducted on "phonological awareness" hardly counts as a convincing counter-argument.

R5. Implicit bias

After reading the commentaries defending implicit bias, there is little reason to believe that implicit bias can help us explain group disparities. A proper response to the target article would have been to outline a causal model for exactly how millisecond differences in simplified judgment tasks lead to group disparities when combined with relevant information in real decision scenarios. The commentaries could have provided a coherent and defensible definition of the concept and showed that measurement of this concept predicted outcomes under the conditions specified by the theory.

None of the commentaries defending implicit bias did this. For example, **Payne and Banaji** chose to not grapple with any evidence or arguments from the target article and instead chose to argue by analogy without doing the necessary work to establish the soundness of the premises foundational to their analogic argument. What features of the physical sciences are responsible for their success? Does implicit bias research have these relevant features to allow for analogic comparisons? What is the evidence that implicit bias researchers have done the work necessary to meet the standards of the other sciences cited in their commentary? Given that research on implicit bias has not answered *even basic definitional questions* despite over two decades of high-volume research activity (Gawronski, 2019; Machery, 2021), the comparison to other sciences is fallacious.

Most commentators, however, insisted that I missed key, convincing studies that demonstrated the importance of implicit bias for understanding group disparities. Let us consider the studies cited by these commentators.

From **Kurdi and Dunham**'s commentary:

1. Hagiwara et al. (2013); Penner et al. (2010); Penner et al. (2016). I start with this trio of papers because as a group they illustrate the problems endemic in citing implicit bias research. **Kurdi and Dunham** cite these papers in support of the claim that "doctors' implicit evaluations predict actual rapport, satisfaction, and treatment adherence among Black patients."

First and most important, **Kurdi and Dunham** incorrectly report the findings of Hagiwara et al. The data unequivocally do not show that "doctors' implicit evaluations predict actual rapport, satisfaction, and treatment adherence among Black patients." The only relevant finding from that paper is that doctors with higher-implicit biases had higher ratios of physician talk time:patient talk time. This study separately analyzed measures of implicit bias and explicit bias and found *no relation* between physicians' implicit or explicit biases and Black patients' adherence or trust of the physician, directly contradicting Kurdi and Dunham's claim.

Penner et al. (2010) tested how physicians' implicit and explicit biases related to Black patients' impressions of them. Working from an aversive racism framework, these authors predicted and found that when physicians had a *combination* of high-implicit bias *plus* low explicit bias, this correlated with negative impressions of physicians by Black patients.

Penner, Dovidio, and colleagues, now in 2016, again studied the relationship between physicians' implicit biases and Black patients' responses. Despite making the case in 2010 that it is the unique *combination* of implicit and explicit bias that matters, here the authors only find a main effect of implicit bias: The greater physicians' implicit biases, the more negative Black patients' impressions of them. What about explicit bias? Although no information about explicit measures was reported in the publication, a measure of explicit bias was included in the study but went unreported because of "no variability" on that measure (personal communication with first author, 6/3/21). It is perfectly reasonable to argue that a lack of variability prevents a meaningful test of a hypothesis; the problem is that the obtained result contradicts the findings of the 2010 paper but is still reported as showing the importance of implicit bias. Moreover, these authors failed to replicate their own findings from Hagiwara et al. concerning the talk time ratio, but no mention of this discrepancy is made.

To consider these three papers at once, then: When high-implicit bias *paired with* low explicit bias leads to negative impressions (Penner et al., 2010), this proves the importance of implicit bias. When high-implicit bias with *no consideration* of explicit bias leads to more negative impressions (Penner et al., 2016), this also proves the importance of implicit bias. Even when high-

https://doi.org/10.1017/S0140525X21000017 Published online by Cambridge University Press

implicit bias produces *no negative effects at all* (Hagiwara et al., 2013), this proves the importance of implicit bias. When direct replications fail (Penner et al., 2016), the original findings still prove the importance of implicit bias. In other words, any pattern of results can be cited in support of implicit bias. Moreover, none of these studies included any non-Black patients, so **Kurdi and Dunham** are simply wrong in claiming that these studies can even speak to group disparities.

As we will see again and again in what follows, the problem with implicit bias generally is the fact that an imprecise and weak theory allows for any outcome to be interpreted as supporting that theory (e.g., Fried, 2020). For example, suppose instead that doctors with high-implicit biases showed *lower* ratios of physician talk time:patient talk time. This also could have been reported as supporting implicit bias, because lower ratios could have been interpreted from an aversive racism framework as doctors not providing enough information to patients, or not being warm enough to patients, or wanting to end the uncomfortable interaction with patients sooner. Similarly, many dependent variables can be interpreted as supportive of implicit bias no matter how they turn out because it is not clear what psychological or social construct they are designed to measure, as the relevant validation work has not been done. What does the ratio of physician talk time:patient talk time actually indicate? Higher values could reflect negativity on the part of the physician, perhaps because the physician talked too much and did not give the patient enough time to talk. On the contrary, higher values could indicate satisfaction by the patient, in that the patient understood the physician and did not feel the need to question her authority. (Indeed, higher ratios correlated with *more adherence* in Hagiwara et al.) In that case, *lower* values could then be interpreted as negativity on the part of the physician, as the reticence of the physician required the patient to ask more questions. Any possible pattern of results can be interpreted as supportive when the concepts and outcomes are poorly defined and validated (see Gelman & Loken, 2013; Kerr, 1998). Below, I provide evidence that this happens in practice as **Kurdi and Dunham** cite *exactly opposite* findings as supporting implicit bias.

Continuing with the work cited by **Kurdi and Dunham**'s commentary:

2. Hehman et al. (2019). This paper was cited by **Kurdi and Dunham** (as well as **Essien et al.** and **Fuentes et al.**) as a convincing illustration of how implicit bias can inform group disparities. This paper reports a correlation between regional levels of implicit bias and racial disparities in fatal police shootings. It is fine to report interesting correlations but the question here is whether the concept of implicit bias adds to our understanding of group disparities. Hehman et al. suggest that citizens' attitudes toward Blacks (as indexed by implicit measures) can "spread" throughout a community via "nonverbal vectors" such as facial expressions. By observing citizens' behaviors toward Blacks, police officers come to "adopt" these same implicit attitudes. Such adopted attitudes can then affect officers' own decision-making, leading them to be more likely to shoot Blacks relative to comparable Whites.

The commentators defending implicit bias cited this work approvingly despite the fact that there is little to no evidence for *any* of the key parts of the model: that implicit bias produces "nonverbal vectors" of the kind proposed; that implicit bias is "contagious" in the manner proposed; or that police officers' decisions to shoot are affected by implicit bias. This lack of evidence is true even for controlled laboratory studies, much less for dynamic neighborhood environments.[1]

Moreover, citing this work illustrates a misunderstanding of the nature of policing and the dynamics of fatal police shootings, which was one of the main points of the target article. A more reasonable interpretation of Hehman et al.'s finding is supported by Johnson and Chopik (2019), cited by none of the commentaries: the missing third variable explaining Hehman et al.'s correlation is actual crime rates.

The same problems apply to other cited works using similar methods, such as that of Riddle and Sinclair (2019) and Chetty et al. (2020).

3. Steffens et al. (2010) is cited by **Kurdi and Dunham** as evidence that implicit bias predicts "actual academic achievement." Steffens et al. find that implicit math self-concept does not predict achievement but implicit stereotypes do. Why a more distal concept should predict the outcome whereas a more proximal concept should not predict the outcome is left unsaid, and certainly nothing about implicit bias predicts this a priori. Given that the effects of self-concept are described as "exploratory," it is possible for any pattern of results to be reported as supportive.

For further illustration of how imprecise theory allows for flexibility in interpreting results, we can return to the cited Penner et al. studies on physicians' implicit bias. Steffens et al. argued that because the implicit association task (IAT) is a comparative measure, a comparative outcome is needed; if math-language implicit associations are measured, then the outcome must be a relative math-language performance measure. But this same logic was not followed by Penner et al., who found that a relative Black–White IAT predicted a single-category outcome, not a relative outcome.

What exactly is measured in studies of implicit bias and how exactly does this relate to the outcomes of interest (see also Blanton, Jaccard, Christie, & Gonzales, 2007)? None of the commentaries made any attempt to offer an answer to this question.

4. Agerstrom and Rooth (2011) is a field audit study and therefore all the problems identified in the target article with these types of studies apply. This work also lacks theory-specific evidence, such as providing evidence that implicit bias effects occur when managers do not want them to happen or that they are unaware of these biases.

5. Glover et al. (2017) is a good study on manager bias and cashier performance, but there is nothing in the study testing any specific, theoretical conditions that would provide unique evidence that the effects are because of implicit bias (as opposed to any number of other variables that might correlate with IAT scores). This also illustrates the flexibility in the patterns of results that can be interpreted as supporting-implicit bias. In Glover et al., both minority and majority managers with high-implicit bias gave *fewer unpleasant work tasks* to minority employees and were *less likely* to ask minority workers to stay late after their shifts; these effects were concentrated in stores with fewer minority workers. Given that this specific pattern was not predicted *a priori*, all possible combinations of manager status, store concentration, dependent variables, and different types of interactions yield an incredible number of possible outcomes that could have been interpreted as supporting-implicit bias (Gelman & Loken, 2013; Kerr, 1998).

https://doi.org/10.1017/S0140525X21000017 Published online by Cambridge University Press

So that the reader does not accuse me of mere hypothetical arguing, here is a definitive, unmistakable example of how *exactly opposite* findings can both be interpreted as support for implicit bias. **Kurdi and Dunham** cite Glover et al.'s finding of being *more* hesitant to talk with minority employees as evidence of managers' implicit bias, while simultaneously citing Hagiwara et al.'s finding of being *less* hesitant to talk to minority patients as evidence of physicians' implicit bias!

6. Olson et al. (2015) is not about disparate outcomes.
7. Dasgupta and Asgari (2004) has no measure of group disparate outcomes.
8. Caliskan and Lewis (2020); Caliskan et al. (2017); Kurdi et al. (2019a); and Charlesworth et al. (2021). All these are cited as showing relationships between implicit bias and "text produced spontaneously and outside any experimental setting." None of these have any measure of group disparities.

Therefore, although **Kurdi and Dunham** claim to "highlight several sets of findings…elucidating the relationship between implicit social cognition and real-world inequality," a closer examination of the cited work shows that there is almost no convincing evidence of this relationship.

From **Mora et al.**'s commentary:

9. Smith and Semin (2007) is cited as relevant to implicit bias and STEM disparities, specifically as it concerns the IAT. This paper makes no mention of STEM and group disparities.
10. Freeman (2014) and Freeman et al. (2016) are both cited in the same sentence as Smith and Semin (2007), yet neither has anything to do with STEM and group disparities.
11. Smeding et al. (2016) is cited as showing "meaningful group differences in decision-making dynamics" but this has no measurement of group disparities.

However, there is something noteworthy about Smeding et al.'s work. **Mora et al.** cite an important aspect of this paper, namely: "Study 3 in Smeding et al. has shown that self-congruency trumps the role of stereotype-congruency in a 'Math v. Language' IAT." Yet recall that Steffens et al. (2010), which was cited by **Kurdi and Dunham**, found effects for stereotype IATs but not self-concept IATs. In other words, when the self-concept trumps stereotypes, one can cite this as support for implicit bias; when stereotypes trump the self-concept, one can also cite this as support for implicit bias. It is only when the details of each study are probed that the inconsistencies and flexibility inherent to this topic are revealed.

12. Shapiro and Williams (2012) is cited as a demonstration of how one's own implicit biases can affect group disparities. This is a summary paper on stereotype threat, and while there is not enough space to adjudicate the debates in the stereotype threat literature, much of this work has not fared well over time (e.g., Flore, Mulder, and Wicherts, 2018). More important for the present purposes, the analysis in the target article still applies to this work. For example, one could ask whether such work incorporates group differences in aptitudes at the tail-ends of the performance distributions.
13. Kutzner and Fiedler (2017) is a summary paper on illusory-like correlations and does not speak directly to group disparities.

R6. Conclusion

The commentaries on the target article ranged from productive comments and descriptions of weaknesses in the target article, to attempts at salvaging the contributions of experimental social psychology generally and implicit bias specifically. These latter attempts mostly failed. Yes, experimental social psychology can do many things, but the contribution of results from such studies to an understanding of group disparities remains unclear. If the goal is to demonstrate that people *can* be biased under certain conditions, this has been thoroughly demonstrated and no sensible person would deny this point. If the goal is to understand why groups obtain different outcomes, the argument in the target article remains intact and methods other than those of traditional social psychology experiments are needed.

Financial support. This article is based on work supported by the National Science Foundation under Grants No. 1230281 and 1756092.

Conflict of interest. None.

Note

1. Hehman et al. cite Weisbuch, Pauker, and Ambady (2009) as evidence that non-verbal behavior transmission can impact implicit biases. In reviewing Weisbuch et al., I found that the key effects reported across the first three study sets, with *N*s of 23, 53, 62, and 35, were: $p = 0.05$, $p = 0.05$, $p = 0.05$, $p = 0.05$, and $p = 0.04$.

References

Agerström, J., & Rooth, D.-O. (2011). The role of automatic obesity stereotypes in real hiring discrimination. *Journal of Applied Psychology*, *96*(4), 790–805. http://doi.org/10.1037/a0021594.

Anderson, K., Ritter, G., & Zamarro, G. (2017). Understanding a vicious cycle: Do out-of-school suspensions impact student test scores?.

Aronson, E. R., & Carlsmith, J. M. (1968). Experimentation in social psychology. In G. Lindzey & E. Aronson (Eds.), *The handbook of social psychology* (Vol. 2, pp. 1–79). Addison-Wesley.

Bauer, M. S., Damschroder, L., Hagedorn, H., Smith, J., & Kilbourne, A. M. (2015). An introduction to implementation science for the non-specialist. *BMC Psychology*, *3*(1), 1–12.

Blanton, H., Jaccard, J., Christie, C., & Gonzales, P. M. (2007). Plausible assumptions, questionable assumptions and post hoc rationalizations: Will the real IAT, please stand up?. *Journal of Experimental Social Psychology*, *43*, 399–409.

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183–186. http://doi.org/10.1126/science.aal4230.

Caliskan, A., & Lewis, M. (2020). Social biases in word embeddings and their relation to human cognition. PsyArXiv. http://doi.org/10.31234/osf.io/d84kg.

Cesario, J. (2021). On selective emphasis, broad agreement, and future directions: Reply to Ross, Winterhalder, & McElreath. https://doi.org/10.31234/osf.io/2p5eg.

Cesario, J., Johnson, D. J., & Terrill, W. (2019). Is there evidence of racial disparity in police use of deadly force? Analyses of officer-involved fatal shootings in 2015–2016. *Social Psychological and Personality Science*, *10*(5), 586–595.

Charlesworth, T. E. S., Yang, V., Mann, T. C., Kurdi, B., & Banaji, M. R. (2021). Gender stereotypes in natural language: Word embeddings show robust consistency across child and adult language corpora of more than 65 million words. *Psychological Science, Psychological Science*, *32*(2), 218–240. http://doi.org/10.1177/0956797620963619.

Chetty, R., Hendren, N., Jones, M. R., & Porter, S. R. (2020). Race and economic opportunity in the United States: An intergenerational perspective. *The Quarterly Journal of Economics*, *135*(2), 711–783. http://doi.org/10.1093/qje/qjz042.

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

Correll, J., Wittenbrink, B., Park, B., Judd, C. M., & Goyle, A. (2011). Dangerous enough: Moderating racial bias with contextual threat cues. *Journal of Experimental Social Psychology 47*:184–189.

Dahl, A. (2017). Ecological commitments: Why developmental science needs 3290 naturalistic methods. *Child Development Perspectives*, *11*(2), 79–84. https://doi.org/329110.1111/cdep.12217.

Darley, J. M., & Gross, P. H. (1983). A hypothesis-confirming bias in labeling effects. *Journal of Personality and Social Psychology*, *44*(1), 20.

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*(5), 642–658. http://doi.org/10.1016/j.jesp.2004.02.003.

Devi, T., & Fryer Jr, R. G. (2020). *Policing the police: The impact of "pattern-or-practice" investigations on crime* (No. w27324). National Bureau of Economic Research.

Dunham, R. G., & Petersen, N. (2017). Making black lives matter: Evidence-based policies for reducing police bias in the use of deadly force. *Criminology & Public Policy*, *16*, 341.

Flore, P. C., Mulder, J., & Wicherts, J. M. (2018). The influence of gender stereotype threat on mathematics test scores of Dutch high school students: A registered report. *Comprehensive Results in Social Psychology*, *3*(2), 140–174.

Freeman, J. B. (2014). Abrupt category shifts during real-time person perception. *Psychonomic Bulletin & Review*, *21*:85–92. https://doi.org/10.3758/s13423-013-0470-8.

Freeman, J., Pauker, K., & Sanchez, D. (2016). A perceptual pathway to bias. *Psychological Science*, *27*(4):502–517. https://doi.org/10.1177/0956797615627418.

Fried, E. I. (2020). Lack of theory building and testing impedes progress in the factor and network literature. *Psychological Inquiry*, *31*(4), 271–288.

Gawronski, B. (2019). Six lessons for a cogent science of implicit bias and its criticism. *Perspectives on Psychological Science*, *14*, 574–595.

Gelman, A., & Loken, E. (2013). The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time. *Department of Statistics, Columbia University*, *348*, 1–17.

Glover, D., Pallais, A., & Pariente, W. (2017). Discrimination as a self-fulfilling prophecy: Evidence from French grocery stores. *The Quarterly Journal of Economics*, *132*(3), 1219–1260. http://doi.org/10.1093/qje/qjx006.

Gordon, D. (2020). The police as place-consolidators: The organizational amplification of urban inequality. *Law & Social Inquiry*, *45*(1), 1–27. http://doi.org/10.1017/lsi.2019.31

Hagiwara, N., Penner, L. A., Gonzalez, R., Eggly, S., Dovidio, J. F., Gaertner, S. L. (2013). 2473 Racial attitudes, physician-patient talk time ratio, and adherence in racially discordant 2474 medical interactions. *Social Science & Medicine*, *87*(C), 123–131. http://doi.org/10.24751016/j.socscimed.2013.03.016.

Hehman, E., Flake, J. K., & Calanchini, J. (2018). Disproportionate use of lethal force in policing is associated with regional racial biases of residents. *Social Psychological and Personality Science 9*:393–401, https://doi.org/10.1177/1948550617711229.

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

IJzerman, H., Lewis, N. A., Przybylski, A. K., Weinstein, N., DeBruine, L., Ritchie, S. J., … Anvari, F. (2020). Use caution when applying behavioural science to policy. *Nature Human Behaviour*, *4*(11), 1092–1094.

Jarvis, S. N., & Okonofua, J. A. (2020). School deferred: When bias affects school leaders. *Social Psychological and Personality Science*, *11*, 492–498.

Johnson, D. J., & Chopik, W. J. (2019). Geographic variation in the black-violence stereotype. *Social Psychological and Personality Science*, *10*(3), 287–294.

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review*, *2*(3), 196–217.

Knox, D., & Mummolo, J. (2020). Making inferences about racial disparities in police violence. *Proceedings of the National Academy of Sciences*, *117*(3), 1261–1262.

Kurdi, B., Mann, T. C., Charlesworth, T. E. S., & Banaji, M. R. (2019a). The relationship between implicit intergroup attitudes and beliefs. *Proceedings of the National Academy of Sciences of the United States of America*, *116*:5862–5871, https://doi.org/10.1073/pnas.1820240116.

Kutzner, F., & Fiedler, K. (2017). Stereotypes as pseudocontingencies. *European Review of Social Psychology*, *28*(1):1–49. https://doi.org/10.1080/10463283.2016.1260238.

Latzer, B. (2018). Subcultures of violence and African American crime rates. *Journal of Criminal Justice*, *54*, 41–49.

Latzer, B. (2020) *Public safety in an era of criminal justice reform*. Manhattan Institute. https://www.youtube.com/watch?v=J8KvHbWSypA.

Lewin, K. (1946). Action research and minority problems. *Journal of Social Issues*, *2*(4), 34–46.

Machery, E. (2021). Anomalies in implicit attitudes research. *Wiley Interdisciplinary Reviews: Cognitive Science*, e1569. https://doi.org/10.1002/wcs.1569

McElreath, R. (2021). *Science before Scientists: Causal Inference*. 2021 Leipzig Spring School in Methods for the Study of Culture and the Mind. https://www.youtube.com/watch?v=KNPYUVmY3NM.

Moss-Racusin, C. A., Dovidio, J. F., Brescoll, V. L., Graham, M. J., & Handelsman, J. (2012). Science faculty's subtle gender biases favor male students. *Proceedings of the National Academy of Sciences*, *109*, 16474–9.

Okonofua, J. A., & Eberhardt, J. L. (2015). Two strikes: Race and the disciplining of young students. *Psychological Science*, *26*, 617–624.

Olson, K. R., Key, A. C., & Eaton, N. R. (2015). Gender cognition in transgender children. *Psychological Science*, *26*(4), 467–474. http://doi.org/10.1177/0956797614568156.

Penner, L. A., Dovidio, J. F., West, T. V., Gaertner, S. L., Albrecht, T. L., Dailey, R. K., & Markova, T. (2010). Aversive racism and medical interactions with black patients: A field study. *Journal of Experimental Social Psychology*, *46*(2), 436–440. http://doi.org/10.1016/j.jesp.2009.11.004.

Penner, L. A., Dovidio, J. F., Gonzalez, R., Albrecht, T. L., Chapman, R., Foster, T. (2016). The effects of oncologist implicit racial bias in racially discordant oncology interactions. *Journal of Clinical Oncology*, *34*(24), 2874–2880. http://doi.org/10.1200/JCO.2015.66.3658.

Premachandra, B., & Lewis Jr, N. (2021). Do we report the information that is necessary to give psychology away? A scoping review of the psychological intervention literature 2000–2018. *Perspectives on Psychological Science*. https://doi.org/10.1177/1745691620974774.

Riddle, T., & Sinclair, S. (2019). Racial disparities in school-based disciplinary actions are associated with county-level rates of racial bias. Proceedings of the National Academy of Sciences of the United States of America, *116*:8255–8260, https://doi.org/10.1073/pnas.1808307116.

Ross, C. T., Winterhalder, B., & McElreath, R. (2021). Racial disparities in police use of deadly force against unarmed individuals persist after appropriately benchmarking shooting data on violent crime rates. *Social Psychological and Personality Science*, *12*, 323–332.

Shapiro, J. R., & Williams, A. M. (2012). The role of stereotype threats in undermining girls' and women's performance and interest in STEM fields. *Sex Roles*, *66*(3–4): 175–183, https://doi.org/10.1007/s11199-011-0051-0.

Skiba, R. J., Chung, C. G., Trachok, M., Baker, T. L., Sheya, A., & Hughes, R. L. (2014). Parsing disciplinary disproportionality: Contributions of infraction, student, and school characteristics to out-of-school suspension and expulsion. *American Educational Research Journal*, *51*(4), 640–670.

Smeding, A., Quinton, J.-C., Lauer, K., Barca, L., & Pezzulo, G. (2016). Tracking and simulating dynamics of implicit stereotypes: A situated social cognition perspective. *Journal of Personality and Social Psychology*, *111*(6):817–834. https://doi.org/10.1037/pspa0000063.

Smith, E. R., & Semin, G. R. (2007). Situated social cognition. *Current Directions in Psychological Science*, *16*(3):132–135. https://doi.org/10.1111/j.1467-8721.2007.00490.x.

Sowell, T. (1983). *The economics and politics of race: An international perspective*. William Morrow and Company.

Sowell, T. (1999). *The quest for cosmic justice*. The Free Press.

Sowell, T. (2015). *Basic economics: A common sense guide to the economy* (5th Ed.). Basic Books.

Sowell, T. (2019). *Discrimination and disparities*. Basic Books.

Steffens, M. C., Jelenec, P., & Noack, P. (2010). On the leaky math pipeline: Comparing implicit math–gender stereotypes and math withdrawal in female and male children and adolescents. *Journal of Educational Psychology*, *102*(4), 947–963. http://doi.org/10.1037/a0019920.

Swencionis, J. K., & Goff, P. A. (2017). The psychological science of racial bias and policing. *Psychology, Public Policy, and Law*, *23*(4), 398.

Weisbuch, M., Pauker, K., & Ambady, N. (2009). The subtle transmission of race bias via televised nonverbal behavior. *Science (New York, N.Y.)*, *326*(5960), 1711–1714.

Welsh, R. O., & Little, S. (2018). The school discipline dilemma: A comprehensive review of disparities and alternative approaches. *Review of Educational Research*, *88*(5), 752–794.

Williams, W. E. (1982). *The state against blacks*. New Press.

https://doi.org/10.1017/S0140525X21000017 Published online by Cambridge University Press