Authors' Response

Reply to the commentaries: A radical revision of experimental social psychology is still needed

Q1 Joseph Cesario

Q2 Michigan State University, East Lansing, MI 48824.
cesario@msu.edu
www.cesariolab.com

doi:10.1017/S0140525X21000000, e0

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, 2020 and Ross, Winterhalder, and McElreath, Q127 in press) 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 & 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 & 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 & 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 US 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.

Hodson also bemoans the shift toward "right-leaning priorities," as if the concerns of roughly half the US 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 ("anyone not on the political right") 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 & Balcetis provide a compelling account of how differences in visual processing may represent an important source of bias. Relatedly, **Oyserman & 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 & Boutwell extend the target article in a useful way to other methods of investigation, including self-report of discrimination experiences. **Salmon & Hehman** build a convincing case for the importance of incorporating ultimate causes in understanding stereotyping processes and outcomes. More proximally, **Rennels & Insouvanh**, coming from a developmental view, connect the target article to a similar critique by Dahl (2017) and **Q128** 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 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). Q129

Duell & 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 US Whites by Japanese-, Chinese-, Indian-, and Caribbean-Americans, to the disconnect between levels of prejudice and the variation in achievement among the US 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 & 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 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 et al. (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 et al. (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 STEM representation." Yet this is precisely the problem discussed in 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 & 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 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's 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 Q130 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's 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.

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 Q131 work." Premachandra and Lewis (2020) 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, **White**, **Kurdi & Dunham**, **Brownstein et al.**, and **Jetten, Selvanathan, Crimston, Bentley, and Haslam** (**Jetten et al.**) 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 & 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 & 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 & 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 & 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 & 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 & Dunham's claim.

Penner, Dividio, and colleagues (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-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 & 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 & Dunham** cite *exactly opposite* findings as supporting-implicit bias.

Continuing with the work cited by **Kurdi & Dunham**'s commentary:

2. *Hehman* et al. *(2019)*. This paper was cited by **Kurdi & 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 & 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, and 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).

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 & 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 & 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 & 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

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.

Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology*, **100** (3), 407. Q132

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.

Cesario, J. (2020). 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.

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. Q133

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

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.

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.

Fried, E. I. (2020). Lack of theory building and testing impedes progress in the factor and network literature. *Psychological Inquiry*, **31**(4), 271–288.

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**. Q134

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.

Johnson, D. J., & Chopik, W. J. (2019). Geographic variation in the black-violence stereotype. *Social Psychological and Personality Science*, **10**(3), 287–294.

Johnson, D. J., & Wilson, J. P. (2019). Racial bias in perceptions of size and strength: The impact of stereotypes and group differences. *Psychological Science*, **30**(4), 553–562. Q135

Jussim, L. (2012). *Social perception and social reality: Why accuracy dominates bias and self-fulfilling prophecy*. OUP USA. Q136

Jussim, L., Cain, T. R., Crawford, J. T., Harber, K., & Cohen, F. (2009). The unbearable accuracy of stereotypes. *Handbook of Prejudice, Stereotyping, and Discrimination*, **199**, 227. Q137

Jussim, L., Crawford, J. T., & Rubinstein, R. S. (2015). Stereotype (in)accuracy in perceptions of groups and individuals. *Current Directions in Psychological Science*, **24**(6), 490–497. Q138

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.

Kunda, Z., & Thagard, P. (1996). Forming impressions from stereotypes, traits, and behaviors: A parallel-constraint-satisfaction theory. *Psychological Review*, **103**(2), 284. Q139

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. Q140

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. Q141

Macrae, C. N., & Bodenhausen, G. V. (2000). Social cognition: Thinking categorically about others. *Annual Review of Psychology*, **51**(1), 93–120. Q142

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.

Premachandra, B., & Lewis Jr, N. (2020). 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. Q115

Ross, C. T., Winterhalder, B., & McElreath, R. (in press). 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*. Q143

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.

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.

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.