RUNNING HEAD: Commentary on Gawronski et al.

Avoiding Bias in the Search for Implicit Bias

Wilson Cyrus-Lai & Warren Tierney
INSEAD

Christilene du Plessis, My Nguyen, Michael Schaerer
Singapore Management University

Elena Clemente
Stockholm School of Economics

Eric Luis Uhlmann
INSEAD

Corresponding authors: wilson-cyrus.lai@insead.edu, eric.luis.uhlmann@gmail.com

Acknowledgments: Wilson Cyrus-Lai, Warren Tierney, & Eric Luis Uhlmann are grateful for R&D support from INSEAD for some of the research discussed here. This commentary benefited from conversations with Wil Cunningham a long time ago.

To revitalize the study of unconscious bias, Gawronski, Ledgerwood, and Eastwick (2022) propose a paradigm shift away from implicit measures of intergroup attitudes and beliefs. Specifically, researchers should capture discriminatory biases and demonstrate that participants are unaware of the influence of social category cues on their judgments and actions. Individual differences in scores on implicit measures will be useful to predict and better understand implicitly prejudiced behaviours, but the latter should be the collective focus of researchers interested in unconscious biases against social groups.

We welcome Gawronski et al.'s (2022) proposal and seek to build on their insights. We begin by summarizing recent empirical challenges to the implicit measurement approach, which has for the last quarter century focused heavily on capturing individual differences and examining their potential antecedents and consequences. In our view, Gawronski et al. underestimate the problems the subfield of implicit bias research is currently facing; the need for a paradigm shift in focus and approach is truly urgent.

Although we strongly agree with their basic thesis, we also stress the importance of avoiding various forms of potential bias in the search for implicit bias. First, research in this area should leverage open science innovations such as pre-registration of competing predictions to allow for intellectually and ideologically dissonant conclusions of equal treatment and "reverse" discrimination against members of historically privileged groups. Second, in assessing awareness of bias, researchers should avoid equating unconsciousness with the null hypothesis that evidence of awareness will not emerge, and instead seek positive evidence that the behavioural bias is implicit in nature. Finally, to avoid underestimating the pervasiveness of intergroup bias, scientists should continue to develop and attempt to validate implicit measures of attitudes and beliefs, which may tap latent prejudices expressed in only a small subset of overt actions.

Empirical challenges to the implicit measurement paradigm

Implicit and indirect measures such as the Implicit Association Test (Greenwald, McGhee, & Schwartz, 1998), evaluative priming (Fazio, Jackson, Dunton, & Williams, 1995), the Affect Misattribution Procedure (Payne, Cheng, Govorun, & Stewart, 2005), and others aim to assess individual differences in intergroup prejudice and stereotypes (for reviews, see Gawronski, De Houwer, & Sherman, 2020; Fazio & Olson, 2003; Uhlmann, Leavitt, Menges, Koopman, Howe, & Johnson, 2012). Such attitudes and beliefs, most often captured as automatic associations, are posited by many scholars to guide judgments and behaviours outside of awareness (e.g., Banaji, Lemm, & Carpenter, 2001; Devine, Forscher, Austin, & Cox, 2012; Greenwald & Krieger 2006; Kang, 2005; Kihlstrom, 2004; cf. Greenwald & Lai, 2021). However, the relationship between scores on implicit measures and relevant outcomes should, at least according to some theories, be moderated by the motivation and ability to engage in effortful correction (Fazio, 1990; Gawronski & Bodenhausen, 2006; cf. Greenwald & Banaji, 2017).

In our view, the once thriving research program on implicit measures of social cognition has lost significant momentum over the last decade due to a set of empirical challenges, a number of which are noted by Gawronski et al. (2022). Perhaps most prominent is progressively less impressive evidence of predictive validity, an apparent decline effect (Schooler, 2011) that could be due to improvements in research practices (Motyl et al., 2017; Nelson, Simmons, & Simonsohn, 2018) as well as intellectual allegiance bias (Berman & Reich, 2010) in some earlier investigations and empirical reviews. Bakker, van Dijk, and Wicherts (2012) report evidence of publication bias in early race IAT predictive validity studies. The most up-to-date meta-analytic results suggests the correlation between individual differences in automatic associations with social groups and relevant judgments and behaviours is positive but weak (*r* = .10, or 1% of the variance in behavioural outcomes;

Kurdi et al., 2019; for earlier meta-analyses, see Cameron, Brown-Iannuzzi, & Payne, 2012; Greenwald, Poehlman, Uhlmann, & Banaji, 2009; Oswald, Mitchell, Blanton, Jaccard, & Tetlock, 2013). Further, the theoretically expected moderators of the controllability of the behaviour and its likelihood of being driven by unconscious factors do not appear to moderate association-behaviour correlations.

Even small implicit discriminatory biases, repeated over many decisions, could accumulate over time causing large inequalities in outcomes between social groups (Greenwald, Banaji, & Nosek, 2015; Hardy et al., 2022). However, this cumulative implicit bias thesis requires high levels of bias on implicit measures (e.g., strong preference for White over Black on the Implicit Association Test) to translate into behavioural discrimination against the target group (e.g., higher probability of selecting White over Black candidates for jobs). Yet re-analyses of at least some published laboratory studies reveal a pattern of pro-Black bias on the outcome measure, with high IAT scores predicting less pro-Black behaviours or equal treatment of Whites and Blacks (Blanton et al., 2009; Schimmack, 2019). This may reflect social desirability bias on some laboratory behavioural measures that leaves the individual-differences correlation between the implicit measure and dependent variable intact. But even if so, this still means simulations of real-world disparities in treatment cannot be readily grounded in aggregated correlational relationships between implicit measures and behaviours; they must also take into account the presence or absence of social category cue effects on outcomes.

Further meta-analytic evidence suggests that the automatic associations tapped by some of the most widely used implicit measures could be causally inert. Forscher et al. (2019) examined studies that manipulated scores on implicit measures (e.g., via an intervention designed to reduce implicit prejudice), and also included behavioural outcomes (e.g., seating distance from a Black or White research confederate). Shifts in associations were unrelated to behavioural change, and did not mediate causal effects of experimental interventions on behaviour. Additional evidence indicates that a successful habit-breaking intervention that reduces biased behaviour in the field is not driven by changes in automatic associations (Forscher et al. 2017). Thus, even if weakly correlated with behavioural outcomes (Kurdi et al., 2019; Oswald et al., 2013), automatic associations could be a mere cognitive residue of past actions and experiences rather than a direct contributor to them (Forscher et al., 2019). The field of implicit social cognition has not sufficiently grappled with the results of this line of research, which questions the long-assumed causal role of automatic associations in human actions.

An alternative perspective is provided by the theory of the bias of crowds (Payne, Vuletich, & Lundberg, 2017), which posits that implicit measures capture cultural level prejudices and stereotypes that most effectively predict aggregate (not individual) level outcomes. Scores on implicit measures are unstable across time within a given individual (Gawronski, Morrison, Phills, & Galdi, 2017), yet reliable across time within communities (Hehman et al., 2019; Payne et al. 2017). A regional history of slavery predicts anti-Black bias on the IAT (Payne, Vuletich, & Brown-Iannuzzi, 2019), and aggregated IAT scores in turn correlate with the use of lethal force by police against Black Americans within a given geography (Hehman, Flake, & Calanchini, 2018). Higher reliabilities and macro-level correlations with variables such as Black vs. White mortality rates, racial disparities in infant health, racially charged internet searches, county-level racial disparities in poverty rates, and national gender gaps in math and science (Hehman, Calanchini, Flake, & Leitner, 2019; Leitner, Hehman, Ayduk, & Mendoza-Denton, 2016; Nosek et al., 2009; Orchard & Price, 2017; Rae et al., 2015) could result in whole or in part from the reduction of measurement error via aggregation (Conner & Evers 2020). They may also be partly due to implicit measures tapping into broader cultural biases with limited implications for individual-level judgments and actions (Arkes & Tetlock, 2004; Mitchell & Tetlock 2006; Olson & Fazio, 2004; Uhlmann, Brescoll, & Paluck, 2006; cf. Nosek & Hansen, 2008).

The above suggests that after a quarter century, the implicit measurement approach to implicit bias has suffered from significant paradigm degeneration (Lakatos, 1970). To maintain itself, auxiliary assumptions such as multiple moderators in conjunction leading to respectable predictive validity correlations (Kurdi et al., 2019), social desirability bias on laboratory behavioural measures (Tierney et al. 2020), the cumulative consequences of minute discriminatory biases (Greenwald et al., 2015; Hardy et al., 2022), mismatched and suboptimal behavioural outcomes in studies examining causality (Gawronski et al., 2022), and aggregate-level crowd biases (Payne et al., 2017) must be invoked. Some or even all these defenses may hold empirically. And yet this heavily modified theoretical structure would still represent a major retreat from earlier models in which pervasive individual-level implicit prejudices and stereotypes constitute major causal contributors to societal inequities. Thus, we believe that Gawronski et al. (2022) underestimate the seriousness of the empirical challenges to the "bias on implicit measures" (BIM) paradigm, as well the need for major reforms including (but not limited to) those they advocate.

Avoiding bias in assessing the prevalence and direction of group-based discrimination

In searching for "unconscious biases people do not know they have" (Gawronski et al., 2022) it makes sense to first identify biased and discriminatory behaviour, and then probe to see if people are aware of being influenced by social category cues. At the same time, especially given criticisms that the implicit bias program is itself biased towards a left-leaning narrative of pervasive prejudice (Arkes & Tetlock, 2004; Mitchell & Tetlock 2006), investigators should build in methodological safeguards that allow us to conclude a lack of behavioural bias or even "reverse" discrimination (i.e., bias against members of high status and positively stereotyped groups).

We can accomplish this by defining our sample space in advance, constraining our analytic flexibility, and pre-committing to publish the research regardless of the outcome. Are we sampling representatively from the domains and outcomes where disparate treatment might emerge? Or specifically selecting contexts where discrimination is more likely, knowingly creating a selection bias? If so, this should be made transparent from the outset. The recent renaissance of methodological reforms in psychology and other sciences (Nelson et al., 2018) offers tools that should limit political bias and further facilitate robust and generalizable conclusions. These include pre-registration of analysis plans (Wagenmakers et al., 2012), registered reports (Chambers et al., 2015; Scheel, Schijen, & Lakens, 2021), direct replications (Simons, 2014), multiverse and crowd analyses (Steegen, Tuerlinckx, Gelman, & Vanpaemel, 2016; Schweinsberg et al., 2020; Silberzahn et al., 2018; Simonsohn, Simmons, & Nelson, 2020), open data to facilitate reanalyses (Simonsohn, 2013), forecasting tournaments (Dreber et al., 2015; Tetlock, Mellers, Rohrbaugh, & Chen, 2014), adversarial collaborations (Clark & Tetlock, 2022; Mellers, Hertwig, & Kahneman, 2001) and crowdsourcing data collections across many locations (Klein et al., 2014; Open Science Collaboration, 2015).

Recently Schaerer et al. (2022) carried out a pre-registered meta-analysis of 87 field audits of gender discrimination conducted in 26 countries over a 44-year time span. To optimize the methods and avoid researcher bias, we employed the innovative red team approach (Lakens, 2020). In parallel to the “blue team” leading the project, an independent “red team” of experts on meta-analysis methods and gender, as well as a librarian, reviewed all aspects of the research plan and provided critical feedback. The meta-analytic results, encompassing 373,706 individual job applications, indicate a statistically significant decline between 1978 and 2021 in discrimination against female applicants for stereotypically male-typed and neutral-typed jobs (e.g., manager, banker, accountant). In contrast, bias in selection against male applicants for stereotypically female-typed jobs (e.g., receptionist, nurse, elementary school teacher) remained stable across the decades. Although no aggregate selection bias against female applicants occurred over the last decade in the nations sampled, we observed very high heterogeneity of effect sizes across different field studies. Such variability is consistent with pro-male behavioural biases in some organizations and contexts, and pro-female behavioural biases in others (see also Kline, Rose, & Walters, 2021). Contemporary pro-male discrimination likely reflects the persistence of some explicit and implicit sexist stereotypes and beliefs (Charlesworth & Banaji, 2022; Eagly, Nater, Miller, Kaufmann, & Sczesny, 2020; Haines, Deaux, & Lofaro, 2016). In contrast, preferences for female applicants for traditionally male jobs (e.g., manager, banker) may be driven by diversity-and-inclusion goals (Chang, Milkman, Chugh, & Akinola, 2019; Leslie, Manchester, & Dahm, 2017; Naumovska, Wernicke, & Zajac, 2020) and resentment of existing power structures and high-status groups (Reynolds, Zhu, Aquino, & Strejcek, 2021).

Any discrimination observed in rigorous future studies could therefore not only be unconscious or conscious (Gawronski et al., 2022), but either consistent with or directly contrary to (i.e., in reaction against) traditional societal stereotypes and prejudices. Social cue based explicit and implicit behavioural biases could be pro-male, pro-female, anti-Black, pro-Black, and so forth (Axt, Ebersole, & Nosek, 2016; Chang et al., 2019; Leslie et al., 2017; Naumovska et al., 2020; Quillian, Pager, Hexel, & Midtbøen, 2017; Reynolds et al., 2020, 2021). Given that most people explicitly endorse equal treatment as a moral ideal (Reynold et al., 2021), behavioural biases favouring members of subordinate groups may often occur automatically (Glaser & Knowles, 2008; Moskowitz, Gollwitzer, Wasel, & Schaal, 1999; Moskowitz & Li, 2011) and even unconsciously (Axt et al., 2016).

To address these important questions more systematically, Schaerer et al. (2022) called for crowdsourced direct replications of influential group-based discrimination paradigms. Two such initiatives focused on gender and racial bias are currently in their initial stages. Notably, older experiments on social cue-based discrimination may fail to emerge in contemporary data collections not only because of advances in research methods (Nelson et al., 2018) but also because of changes in the broader society (i.e., cultural evolution, Varnum & Grossmann, 2017). Thus, revisiting influential experimental demonstrations of discriminatory behaviour represents a critical early step in the search for implicit bias. For example, consistent with their aversive racism model of subtle and rationalized implicit prejudice, Dovidio and Gaertner (2000) observed preferences for White over Black job applicants only when job qualifications were ambiguous. In another widely cited investigation, Gawronski, Geschke, and Banse (2003) demonstrated that ambiguous behavioural descriptions were interpreted significantly more negatively for Turkish targets than for German targets, and that scores on a German-Turkish attitudes IAT predicted such biased impressions. Would these main effects of target race and ethnicity replicate in the 2020s? Would awareness tests suggest the influence of social cues was unconscious in nature? And would individual differences in automatic associations still predict the behavioural biases in studies such as those by Gawronski et al. (2003), and extend to further experimental designs such as the Dovidio and Gaertner (2000) aversive racism in hiring paradigm? Large-scale replication methods are best positioned to answer these questions, and to prevent researcher bias towards any specific answer. New data collections should further engage in conceptual replications (Simons, 2014), optimizing designs based on expert feedback (Vohs et al., 2021) and adding further measures and conditions facilitating competitive theory-testing (Tierney et al., 2020).

Recent efforts to self-replicate previously published discrimination effects from the present last author and his collaborators might (and might not) foreshadow the results of broader initiatives to come. Gawronski et al. (2022) cite Uhlmann and Cohen’s (2005, 2007) investigations of constructed criteria and illusions of objectivity in selection decisions, highlighting how such processes may contribute to implicit behavioural bias (see also Hodson, Dovidio, & Gaertner, 2002; Norton, Vandello, & Darley, 2004, for similar results). Tierney et al. (2020) recently conducted a large-sample self-replication that validated these processes but inverted the direction of the social cue effect. In a mirror image of the results from Uhlmann and Cohen (2005, 2007), participants constructed criteria biased against male candidates for the job of police chief and engaged in greater discrimination against men when led to feel objective. Individuals who strongly rejected sexism and had more experience with research studies were especially likely to select a woman for a stereotypically male-typed role, consistent with an inclusion motives and shift in public norms account.

We also recently completed a large-scale crowdsourced initiative re-examining the relationships between workplace emotion expression, the gender of person who expresses the emotion, and how social perceivers evaluate that person. This follows on experimental studies conducted approximately two decades ago and published some years later (Brescoll & Uhlmann, 2008), finding backlash effects against angry women in terms of their perceived competence as well as the degree of social status and respect they receive. Prior work points to the implicit roots of such prescriptive stereotype effects (Rudman & Glick, 2001). Two recent multi-national replication studies collected over eleven thousand participants from more than 20 nations who were assigned to 27 different conceptual replication designs (Tierney et al., 2022). Overall, we find that expressing anger increases status by boosting perceived assertiveness and dominance, and at the same time reduces status by diminishing competence and likability. The downstream consequences of expressing anger vs. sadness or neutral emotion were similar for both female and male targets, across nations, in adult and student samples, and among female and male social perceivers. We therefore failed to replicate the original Brescoll and Uhlmann (2008) findings of bias against angry women, potentially due to shifts in norms related to gender in the intervening time period (Schaerer et al., 2022) and perhaps also cultural changes in the social signals sent by becoming angry in work settings.

Forecasting data indicate such results are highly unexpected to academics. When asked to predict the results of Tierney et al. (2020) based on the materials and methods alone, independent scientists were remarkably accurate overall despite the complex design and interaction tests involved. The glaring exception was the main effect of target gender, which the crowd of forecasters predicted in precisely the wrong direction. Scientists expected the original Uhlmann and Cohen (2005) pattern of bias against female job candidates to emerge again nearly two decades later, yet the large-sample replication revealed directly contrary results. Academic forecasters similarly expected that the original backlash effect against angry women (Brescoll & Uhlmann, 2008) would replicate, that female targets would be conferred less status than male targets overall, and that recent field audits would reveal selection biases against female candidates for stereotypically male-typed and neutral-typed jobs (Schaerer et al., 2022; Tierney et al., 2022). Such strong priors could create ideological blind spots for investigators (Arkes & Tetlock, 2004; Mitchell & Tetlock, 2006), which we argue can be counteracted via open science best practices.

Avoiding bias in attributions of consciousness vs. unconsciousness

Once a discriminatory bias (in either direction) is established, the next challenge is to determine whether social perceivers are aware of the causal influence of the social category cue. This returns us to a longstanding controversy in the literature on unconscious cognition, including subliminal perception (Draine & Greenwald, 1998; Holender, 1986), unconscious learning (Eriksen, 1960; Shanks, Malejka, & Vadillo, 2021), and introspection into mental processes (Ericsson & Simon, 1980; Nisbett & Wilson, 1977). Specifically, by what criteria do we distinguish consciousness from unconsciousness?

Methodologically, the standard approach is to include measures of conscious awareness towards the end of the experiment, and if participants fail to report any such awareness conclude that the underlying psychological processes were unconscious (Bargh & Chartrand, 2000). This creates the “problem of the null” (Uhlmann, 2014), in that unconsciousness becomes the null hypothesis that significant evidence of awareness will not emerge. This sets a lax criterion for unconsciousness in that forgetfulness, asking the wrong probe questions, and measurement error are potentially conflated with a lack of awareness (Shanks et al., 2021; Uhlmann, Pizarro, & Bloom, 2008). In the domain of implicit behavioural bias, self-report measures of awareness are further compromised by social desirability concerns: decision makers may be reluctant to openly admit to discriminating based on race, gender, and other morally charged target characteristics.

Gawronski et al. (2022) propose to therefore rely on experimental paradigms in which decision makers are both 1) motivated to be unbiased and 2) able to consciously control their responses. If such conditions can be assured, any behavioural bias that emerges is likely to be unconscious in nature. Although it is easy to identify tasks where responses are at least in principle controllable (e.g., hiring decisions made without time pressure), ensuring that participants are genuinely motivated to be unbiased again raises concerns about socially desirable responding. Participants could falsely report wanting to treat others equally, and yet engage in covert discrimination on behavioural measures where bias can be detected in the aggregate but not at the level of individual decision makers (see Kuklinski, Cobb, & Gilens, 1997). Incentivizing more accurate and unbiased responding, for example with financial payoffs (Axt et al., 2016), risks equating a manipulation failure with unconsciousness, running once again into the problem of the null.

There exists no perfect awareness criterion, only those with different costs and benefits and that vary in how liberal and conservative they are in inferring consciousness and the lack thereof. Is it the investigators' goal to provide strong and conclusive evidence, or weak and initial evidence, of the implicit nature of the bias? If initial evidence, a robust and replicable discrimination effect and little to no indication of awareness on funnelled debriefing questions at the end of the experiment (Bargh & Chartrand, 2000) are sufficient. But to make a strong claim of implicit behavioural bias, a more conservative test offering positive evidence of unconsciousness is needed (Uhlmann, 2014; Uhlmann et al., 2008).

Drawing on the literature on prime-to-behaviour effects (Bargh & Chartrand, 1999), one potential tactic is to add an experimental condition further increasing the salience of the manipulated variable (for examples see Erb, Bioy, & Hilton, 2002; Martin, Seta, & Crelia, 1990; Moskowitz & Roman, 1992; Moskowitz & Skurnick 1999; Newman & Uleman, 1990; Strack, Schwarz, Bless, Kübler, & Wänke, 1993). If the influence of the social category cue (e.g., race) is eliminated or reversed under conditions that promote greater attention and awareness, this suggests that the discrimination in the low-cue-salience condition occurred unconsciously. For example, Dovidio and Gaertner (2000) manipulated candidate race with a relatively subtle cue, specifically membership in either the Black Student Union or a historically majority-White fraternity. If racial category membership were to be activated more blatantly and repeatedly, the anti-Black discrimination effect might vanish or reverse even in the ambiguous qualifications condition. Contrarily, if decision makers are consciously biased against a target group, discrimination should remain constant or even increase when group membership is made more cognitively accessible. A related approach is to manipulate whether targets are evaluated jointly or separately (Bohnet, van Geen, & Bazerman, 2012). Behavioural discrimination in a between-subjects comparison, which is eliminated or reversed in a within-subjects comparison, suggests the former occurs outside of awareness or is at the very least counteracted by enhanced awareness and detectability (Bohnet et al., 2012; Kuklinski et al., 1997).

Similar inferences can be drawn from a significant interaction between scores on a funnelled debriefing (Bargh & Chartrand, 2000) and the manipulation of target group membership. If participants who express no suspicion of being influenced by the experimental manipulation exhibit the hypothesized effect, but suspicious participants do not, the causal influence among the non-suspicious was probably unconscious (Lombardi, Higgins, & Bargh, 1987; Newman & Uleman, 1990). Such an interaction pattern also validates the awareness measure, eliminating at least one counter-explanation for apparent unconsciousness of being influenced. If responses on the awareness probe reliably moderate the effects of the experimental manipulation, the probe questions are sufficiently relevant, sensitive, and immediate to capture awareness.

As Bargh and Hassin (2022) caution, we should not make conscious awareness the default conclusion either. In most future experiments on behavioural discrimination, neither a high standard for inferring consciousness nor unconsciousness of the influence of the social category cue will be met. Another pragmatic concern is that rigorously measuring and manipulating awareness is much easier in the controlled environs of the laboratory, and yet behavioural discrimination against low status and negatively stereotyped groups is far more common in field settings. Contrast the laboratory results of Axt et al. (2016) who observe a replicable pro-Black bias in judgments that meets meaningful criteria for unconsciousness (Bargh & Chartrand, 2000; Gawronski et al., 2022), with the Quillian et al. (2017) meta-analysis of field audits revealing systematic anti-Black bias in actual selection decisions (see also Kline et al., 2021). The question then arises what the limited ability to make strong claims of unconsciousness in field settings, or readily capture real-world discriminatory tendencies in the laboratory, means for a science of implicit bias that has shifted its focus to behaviour.

Implicit measures could tap latent bias and behavioural measures expressed bias

We agree with Gawronski et al. (2022) that bias on implicit measures (BIM) is a potential indicator of implicit behavioural bias (IB) and a tool with which to better understand it. At the same time, considering the results of our recent open science investigations of discrimination (Schaerer et al., 2022; Tierney et al., 2020), we believe bias on implicit measures is important to focus on in-and-of itself. Human behaviours are multiply determined, such as by both culturally socialized stereotypes (Banaji et al., 2001; Charlesworth & Banaji, 2022) and contravening forces such as diversity and inclusion motives (Crandall & Eshleman, 2003; Leslie et al., 2017; Fazio, 1990; Reynolds et al., 2021). Because of this, behavioural measures are unlikely to ever represent process-pure reflections of implicit bias (Conrey, Sherman, Gawronski, Hugenberg, & Groom, 2005; Jacoby, 1991; Mayerl, Alexandrowicz, & Gula, 2019). It is therefore valuable to distinguish between a *latent bias* in the individual and *expressed bias* in behavioural outcomes (see Crandall & Eshleman, 2003). Implicit and indirect measures aim to tap a latent underlying bias that may manifest itself in only a small subset of overt actions that are simultaneously driven by other factors as well.

A key piece of Gawronski et al.'s (2022) case against a focus on BIM is that implicit measures do not appropriately capture attitudes that reside entirely outside of conscious awareness. Strong within-subject correlations of .50 or even higher between self-perceived automatic preferences and IAT scores (Hahn, Judd, Hirsh, & Blair, 2014) indicate the relevant associations are automatic, unintentional, efficient, and effortless, yet not unconscious (see also Cunningham, Nezlek, & Banaji, 2004; Cunningham, Preacher, & Banaji, 2001; Ranganath, Smith, & Nosek, 2008; Smith & Nosek, 2011). To a substantial degree, people can sense internal spontaneous reactions, including those that depart from their deliberatively endorsed evaluations (Gawronski & Bodenhausen, 2006; Fazio & Olson, 2003). But if the case for the implicit nature of automatic associations was overstated, the case against the validity of such associations as measures of attitudes and beliefs was overstated as well. In other words, strong individual-level correspondence between self-perceived automatic preferences and implicit measures provide evidence that the latter are valid indicators of such preferences. This is true even absent sizeable correlations with behaviours (Kurdi et al., 2019). It may be the nature of contemporary prejudice for many well-intentioned individuals to internally experience biased thoughts and inferences they are at least partially aware of and must constantly correct for to avoid mistreating others (Devine, Monteith, Zuwerink, & Elliot, 1991).

Implicit measures are also valuable in assessing general evaluative and trait associations (e.g., between the categories “women” and “family”, “men” and “career”, or “African-American” and “Bad”), in contrast to behavioural measures which are specific to a situation and outcome (Ajzen, 1985; Ajzen & Fishbein, 1977). That evaluators in a number of developed countries no longer appear to engage in systematic biases in selection against female job applicants for many jobs (Schaerer et al., 2022) does not mean they are not biased and sexist against women in other ways, for example when it comes to promotions (Goldin, Kerr, Olivetti, & Barth, 2017), wage allocations (Auspurg, Hinz, & Sauer, 2017; Bar-Haim et al., 2018; Joshi, Son, & Roh, 2015), career penalties for parenthood (Dias, Chance, & Buchanan, 2020), sexual harassment (Quick & McFadyen, 2017), or even just their spontaneous thoughts and feelings (Devine et al., 1991). Focusing too much on specific behavioural outcomes, and not enough on the general attitudes, beliefs, and associations individuals hold in their minds, could introduce a different type of bias by systematically underestimating the pervasiveness of culturally socialized prejudices.

At the same time, the extent to which latent automatic biases correlate with micro-level judgments and behaviours remains an important and not yet fully resolved empirical question. It will be incredibly valuable to conduct pre-registered replications of key implicit measure behavioural validation studies— carefully selecting experimental paradigms, contexts, and populations where implicit bias should theoretically emerge and implicit measures ought to exhibit predictive validity. Facilitating this, Kurdi et al. (2019) identify studies characterized by much stronger relations between automatic associations (as measured by the IAT) and criterion measures. These include studies that used difference score measures of behaviour, measured polarized attributes, focused primarily on automatic associations and behaviour, and where the predictor and outcome measures were carefully matched. Drawing on Gawronski et al. (2022), we propose adding the replication selection criteria of overall bias against the minority or underrepresented group on the behavioural outcome measure (e.g., Gawronski et al., 2003). There is no need to choose—we can (re)examine both implicitly biased behaviour (IB) and bias on implicit measures (BIM) together.

A longitudinal approach administering implicit, explicit, and behavioural measures at multiple time points could shed fresh light on the causality issue raised by Forscher et al. (2019). Even if the incremental predictive validity of automatic associations beyond explicit measures is modest (Greenwald et al., 2009; Kurdi et al., 2019; Oswald et al., 2013), there could be indirect effects of automatic associations on behavioural bias via changes in explicit attitudes (Gawronski & Bodenhausen, 2006; Smith, Ratliff, & Nosek, 2012). For example, cultural associations with Black Americans conditioned earlier may lay part of the foundation for more complex explicit beliefs and ideologies that exert both conscious and unconscious influences on discrimination (see Galdi, Arcuri, & Gawronski, 2008, for an analogous result in the domain of political voting). Alternatively, mental associations could reflect the automatization of explicit attitudes, potentially mediating their unconscious influences on behavioural biases. If the cognitive residue hypothesis (Forscher et al., 2019) holds, automatic associations should reflect past behaviours and explicitly endorsed attitudes and fail to independently predict future discrimination above-and-beyond such variables. Longitudinal work could also reveal a dynamic interplay between automatic and explicit attitudes and behaviours, such that these all shape one another through processes of socialization, automatization, and rationalization.

Summary and Conclusion

Gawronski et al.'s (2022) target article promises to revitalize the study of implicit bias via a new collective focus on how social category cues unconsciously influence discriminatory behaviour. Both as researchers and as citizens, we should be primarily concerned with unfair and immoral disparate treatment of social groups in hiring, policing, and other high-stakes outcomes. Although this paradigm shift will be most welcome, we highlight the importance of avoiding bias in the search for implicit bias.

In testing for behavioural discrimination, it will be important to define the sample space in advance. What are the key domains in which discrimination might occur? In which of these contexts is latent implicit bias theoretically expected to express itself in overt behaviour? Emerging best practices of open science such as pre-registering competing predictions (Tierney et al., 2020; Wagenmakers et al., 2012), registered reports (Chambers et al., 2015), red teams (Lakens, 2020), and adversarial collaborations (Clark & Tetlock, 2022) will allow us to better evaluate not only discriminatory bias but also non-bias and "reverse" biases (i.e., instances of better treatment of members of historically disadvantaged groups). Only once we confirm the existence of a bias and ascertain its direction can we probe to see if decision makers are aware of being influenced by social category cues. In doing so, we should set *a priori* criteria for unconsciousness and consciousness that avoid biasing conclusions in either direction, or are at least transparent about whether a lax or strict criterion is being applied. In the long-term, we believe implicit measures will hold continuing value – not only in helping to explain (small) slices of the variance in behavioural discrimination, but also by capturing latent biases that may or may not find expression in a given judgment or action. To properly test this latent bias thesis, future investigations should leverage experimental interventions (Forscher et al., 2019) and longitudinal designs (Galdi et al., 2008) to assess whether automatic associations make any causal contribution to implicit behavioural biases.

If our own recent experiences are any guide, combining a renewed focus on implicit behavioural bias (Gawronski et al., 2022) with the ongoing renaissance in research practices (Nelson et al., 2018) will produce results that deeply challenge our intellectual and ideological commitments. We may not find what we came looking for.

References

Ajzen, I. (1985). From intentions to actions: A theory of planned behavior. In J. Kuhl & J. Beckmann (Eds.), *Action Control: From Cognition to Behavior* (pp. 11–38). Berlin: Springer-Verlag.

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

Arkes, H. R., & Tetlock, P. E. (2004). Attributions of implicit prejudice, or “Would Jesse Jackson ‘fail’ the Implicit Association Test?” *Psychological Inquiry, 15,* 257-279.

Auspurg, K., Hinz, T., & Sauer, C. (2017). Why should women get less? Evidence on the gender pay gap from multifactorial survey experiments. *American Sociological Review*, *82*(1), 179-210.

Axt, J.R., Ebersole, C.R. & Nosek, B.A. (2016). An unintentional, robust, and replicable pro-Black bias in social judgment. *Social Cognition, 34,* 1-39.

Bakker, M., van Dijk, A. & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science, 7*(6), 543-554.

Banaji, M. R., Lemm, K. M., & Carpenter, S. J. (2001). The social unconscious. In A. Tesser & N. Schwartz (Eds.), *Blackwell handbook of social psychology: Intraindividual processes* (pp. 134-158). Oxford, UK: John Wiley & Sons.

Bargh, J.A., & Chartrand, T.L. (1999). The unbearable automaticity of being. *American Psychologist, 54,* 462–479.

Bargh, J. A., & Chartrand, T. L. (2000). A practical guide to priming and automaticity research. In H. Reis & C. Judd (Eds.), *Handbook of research methods in social psychology* (pp. 253-285). New York: Cambridge University Press.

Bargh, J. A., & Hassin, R. (2022). The human unconscious in situ: The kind of awareness that really matters. A. Reber, & R. Allen, *The cognitive unconscious*. New York: Oxford University Press.

Bar-Haim, E., Chauvel, L., Gornick, J. C., & Hartung, A. (2018). The persistence of the gender earnings gap: Cohort trends and the role of education in twelve countries. *LIS Working Paper Series*.

Berman, J. S. & Reich, C. M. (2010). Investigator allegiance and the evaluation of psychotherapy outcome research. *European Journal of Psychotherapy and Counselling*, *12*, 11–21.

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*(3), 567-582.

Bohnet, I., van Geen, A., & Bazerman, M. (2012). When performance trumps gender bias: Joint versus separate evaluation. *Management Science*, *62*(5), 1225–1234.

Brescoll, V., & Uhlmann, E.L. (2008). Can angry women get ahead? Status conferral, gender, and workplace emotion expression. *Psychological Science*, *19*, 268-275.

Cameron, C. D., Brown-Iannuzzi, J. L., & Payne, B. K. (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.

Chambers, C. D., Dienes, Z., McIntosh, R. D., Rotshtein, P., & Willmes, K. (2015). Registered reports: realigning incentives in scientific publishing. *Cortex*, *66*, A1-A2.

Chang, E. H., Milkman, K. L., Chugh, D., & Akinola, M. (2019). Diversity thresholds: How social norms, visibility, and scrutiny relate to group composition. *Academy of Management Journal*, *62*(1), 144-171.

Charlesworth, T. E. S., & Banaji, M. R. (2022). Patterns of implicit and explicit stereotypes

III. Long-term change in gender-science and gender-career stereotypes. *Social Psychological and Personality Science, 13*(1), 14-26.

Clark, C. J. & Tetlock, P. E. (2022). Adversarial collaboration: The next science reform. In C. Frisby, R. Redding, W. O'Donohue, & S. Lilienfeld (Eds.), *Political Bias in Psychology: Nature, Scope, and Solutions*. New York, NY: Springer.

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

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

Crandall, C. S., & Eshleman, A. (2003). A justification–suppression model of the expression and experience of prejudice. *Psychological Bulletin, 129*, 414–446.

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

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

Devine, P. G., Forscher, P. S., Austin, A. J., & Cox, W. T. (2012). Long-term reduction in implicit race bias: A prejudice habit-breaking intervention. *Journal of Experimental Social Psychology, 48,* 1267-1278.

Devine, P. G., Monteith, M. J., Zuwerink, J. R., & Elliot, A. J. (1991). Prejudice with and without compunction. *Journal of Personality and Social Psychology, 60*, 817-830

Dias, F. A., Chance, J., & Buchanan, A. (2020). The motherhood penalty and the fatherhood premium in employment during covid-19: Evidence from the United States. *Research in Social Stratification and Mobility, 69,* 100542.

Dovidio, J. F., & Gaertner, S. L. (2000). Aversive racism and selection decisions: 1989 and 1999. *Psychological Science*, *11*, 319-323.

Draine, S.C., & Greenwald, A.G. (1998). Replicable unconscious semantic priming. *Journal of Experimental Psychology: General, 127*, 286–303.

Dreber, A., Pfeiffer, T., Almenberg, J., Isaksson, S., Wilson, B., Chen, Y., Nosek, B. A., & Johannesson, M. (2015). Using prediction markets to estimate the reproducibility of scientific research. *Proceedings of the National Academy of Sciences, 112*(50), 15343-15347.

Eagly, A., Nater, C., Miller, D., Kaufmann, M., & Sczesny, S. (2020). Gender stereotypes have changed: A cross-temporal meta-analysis of US public opinion polls from 1946 to 2018. *American Psychologist, 75*(3), 301-315.

Erb, H., Bioy, A., & Hilton, D.J. (2002). Choice preferences without inferences: Subconscious priming of risk attitudes. *Journal of Behavioral Decision Making, 15,* 251–262.

Ericsson, K., & Simon, H. (1980). Verbal reports as data. *Psychological Review, 87*(3), 215-251.

Eriksen, C. W. (1960). Discrimination and learning without awareness: A methodological survey and evaluation. *Psychological Review, 67*(5), 279–300.

Fazio, R. H. (1990). Multiple processes by which attitudes guide behavior: The MODE model as an integrative framework. *Advances in Experimental Social Psychology, 23,* 75-109.

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

Fazio, R. H., & Olson, M. A. (2003). Implicit measures in social cognition research: Their meaning and use. *Annual Review of Psychology, 54,* 297-327.

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,* 522-559.

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

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

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

Gawronski, B., De Houwer, J., & Sherman. J. W. (2020). Twenty-five years of research using implicit measures. *Social Cognition, 38*, s1-s25.

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

Gawronski, B., Ledgerwood, A., & Eastwick, P.W. (2022). Implicit bias ≠ bias on implicit measures. (Target Article). *Psychological Inquiry.*

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

Glaser, J., & Knowles, E. D. (2008). Implicit motivation to control prejudice. *Journal of Experimental Social Psychology, 44*(1), 164-172.

Goldin, C., Kerr, S. P., Olivetti, C., & Barth, E. (2017). The expanding gender earnings gap: Evidence from the LEHD-2000 Census. *American Economic Review, 107*(5), 110-114.

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

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

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

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

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

Greenwald, A. G., Poehlman, T. A., Uhlmann, E. 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*(1), 17.

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

Haines, E. L., Deaux, K., & Lofaro, N. (2016). The times they are a-changing… or are they not? A comparison of gender stereotypes, 1983–2014. *Psychology of Women Quarterly, 40*(3), 353-363.

Hardy III, J. H., Tey, K. S., Cyrus-Lai, W., Martell, R. F., Olstad, A., & Uhlmann, E. L. (2022). Bias in context: Small biases in hiring evaluations have big consequences. *Journal of Management*, *48*(3), 657-692.

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*, 1022-1040.

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.

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

Holender, D. (1986). Semantic activation without conscious identification in dichotic listening, parafoveal vision, and visual masking: A survey and appraisal. *Behavioral and Brain Sciences, 9*, 1–23.

Jacoby, L. L. (1991). A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory and Language*, *30,* 513-541.

Joshi, A., Son, J., & Roh, H. (2015). When can women close the gap? A meta-analytic test of sex differences in performance and rewards. *Academy of Management Journal*, *58*(5), 1516-1545.

Kang, J. (2005). Trojan horses of race. *Harvard Law Review, 118*, 1489 – 1593.

Kihlstrom, J. F. (2004). Implicit methods in social psychology. In C. Sansone, C. C. Morf, & A. T. Panter (Eds.), *The Sage Handbook of Methods in Social Psychology* (pp. 195–212). Thousand Oaks, CA: Sage.

Klein, R. A., Ratliff, K. A., Vianello, M., Adams Jr, R. B., Bahnik, S., Bernstein, M. J., Bocian, K., Brandt, M. J., Brooks, B., & Brumbaugh, C. C. (2014). Theory building through replication: Response to commentaries on the" Many labs" replication project. *Social Psychology, 45*(4), 307-310.

Kline, P.M., Rose, E.K., & Walters, C.R. (2021). *Systematic discrimination among large U.S. employers*. Unpublished manuscript, National Bureau of Economic Research.

Kuklinski, J. H., Cobb, M. D., & Gilens, M. (1997). Racial attitudes and the “New South.” *Journal of Politics, 59*, 323–349.

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

Lakatos, I. (1970). Falsification and the methodology of scientific research programmes. In I. Lakatos & A. Musgrave, Eds. *Criticism and the Growth of Knowledge* (pp. 91–195). Cambridge: Cambridge University Press.

Lakens, D. (2020). Pandemic researchers--recruit your own best critics. *Nature, 581*(7807), 121-122.

Leitner, J.B., Hehman, E., Ayduk, O., & Mendoza-Denton, R. (2016). Racial bias is associated with ingroup death rate for Blacks and Whites: Insights from Project Implicit. *Social Science & Medicine, 170,* 220– 227

Leslie, L. M., Manchester, C. F., & Dahm, P. C. (2017). Why and when does the gender gap reverse? Diversity goals and the pay premium for high potential women. *Academy of Management Journal*, *60*(2), 402-432.

Lombardi, W.J., Higgins, E.T., & Bargh, J.A. (1987). The role of consciousness in priming effects on categorization: Assimilation versus contrast as a function of awareness of the priming task. *Personality and Social Psychology Bulletin, 13,* 411–429.

Martin, L.L., Seta, J.J., & Crelia, R.A. (1990). Assimilation and contrast as a function of people's willingness and ability to expend effort in forming an impression. *Journal of Personality and Social Psychology, 59*, 27–37.

Mayerl, H., Alexandrowicz, R.W., & Gula, B. (2019). Modeling effects of newspaper articles on stereotype accessibility in the shooter task. *Social Cognition, 37*, 571–595.

Mellers, B., Hertwig, R., & Kahneman, D. (2001). Do frequency representations eliminate conjunction effects? An exercise in adversarial collaboration. *Psychological Science, 12*(4), 269-275.

Mitchell, G., & Tetlock, P. (2006). Antidiscrimination law and the perils of mindreading. *Ohio State Law Journal, 67*, 1023-1121.

Moskowitz, G.B., Gollwitzer, P.M., Wasel, W., & Schaal, B. (1999). Preconscious control of stereotype activation through chronic egalitarian goals. *Journal of Personality and Social Psychology*, 77, 167-184.

Moskowitz, G. B., & Li, P. (2011). Egalitarian goals trigger stereotype inhibition: A proactive form of stereotype control. *Journal of Experimental Social Psychology, 47*(1), 103-116.

Moskowitz, G.B., & Roman, R.J. (1992). Spontaneous trait inferences as self-generated primes: Implications for conscious social judgment. *Journal of Personality and Social Psychology, 62,* 728–738.

Moskowitz, G.B., & Skurnick, I.W. (1999). Contrast effects as determined by the type of prime: Trait versus exemplar primes initiate processing strategies that differ in how accessible constructs are used. *Journal of Personality and Social Psychology, 6,* 911–927.

Motyl, M., Demos, A. P., Carsel, T. S., Hanson, B. E., Melton, Z. J., Mueller, A. B., Prims, J.

P., Sun, J., Washburn, A. N., Wong, K., Yantis, C. A., & Skitka, L. J. (2017). The state of social and personality science: Rotten to the core, not so bad, getting better, or getting worse? *Journal of Personality and Social Psychology, 113*, 34-59.

Naumovska, I., Wernicke, G., & Zajac, E. J. (2020). Last to come and last to go? The complex role of gender and ethnicity in the reputational penalties for directors linked to corporate fraud. *Academy of Management Journal*, *63*(3), 881-902.

Nelson, L., Simmons, J., & Simonsohn, U. (2018). Psychology's Renaissance. *Annual Review of Psychology, 69*, 511-534.

Newman, L.S., & Uleman, J.S. (1990). Assimilation and contrast effects in spontaneous trait inference. *Personality and Social Psychology Bulletin, 16*, 224–240.

Nisbett, R.E., & Wilson, T.D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review, 84*, 231–259.

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

Nosek, B. A., & Hansen, J. J. (2008). The associations in our heads belong to us: Searching for attitudes and knowledge in implicit evaluation. *Cognition and Emotion*, *22*(4), 553–594.

Nosek, B. A., Smyth, F. L., Sriram, N., Linder, 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.

Olson, M. A., & Fazio, R. H. (2004). Reducing the influence of extrapersonal associations on the Implicit Association Test: Personalizing the IAT. *Journal of Personality and Social Psychology, 86*, 653-667.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), 943.

Orchard, J., & Price, J. (2017). County-level racial prejudice and the Black-White gap in infant health outcomes. *Social Science & Medicine, 181*, 191–198.

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

Payne, B. K., Cheng, S. M., Govorun, O., & Stewart, B. D. (2005). An inkblot for attitudes: Affect misattribution as implicit measurement. *Journal of Personality and Social Psychology, 89*, 277-293.

Payne, B. K., Vuletich, H. A., & Brown-Iannuzzi, J. L. (2019). Historical roots of implicit bias in slavery. *Proceedings of the National Academy of Sciences,* 201818816.

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.

Quick, J. C. & McFadyen, M. A. (2017). Sexual harassment: Have we made any progress? *Journal of Occupational Health Psychology*, *22*(3), 286-298.

Quillian, L., Pager, D., Hexel, O., & Midtbøen, 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.

Rae, J. R., Newheiser, K., & Olson, K. R. (2015). Exposure to racial out-groups and implicit race bias in the United States. *Social Psychological & Personality Science, 6,* 535–543.

Ranganath, K. A., Smith, C. T., & Nosek, B. A. (2008). Distinguishing automatic and controlled components of attitudes from indirect and direct measurement. *Journal of Experimental Social Psychology, 44*, 386-396.

Reynolds, T., Howard, C., Sjastad, H., Okimoto, T., Baumeister, R. F., Aquino, K., & Kim, J. (2020). Man up and take it: Gender bias in moral typecasting. *Organizational Behavior and Human Decision Processes, 161*, 120-141.

Reynolds, T., Zhu, L., Aquino, K., & Strejcek, B. (2021). Dual pathways to bias: Evaluators' ideology and ressentiment independently predict racial discrimination in hiring contexts. *Journal of Applied Psychology, 106*(4), 624–641.

Rudman, L. A., & Glick, P. (2001). Prescriptive gender stereotypes and backlash toward agentic women. *Journal of Social Issues, 57*, 743–762.

Schaerer, M., du Plessis, C., Nguyen, M., van Aert, R.C.M., Tiokhin, L., Lakens, D., Clemente, E., Pfeiffer, T., Dreber, A., Magnus Johannesson, M., Clark, C.J., Gender Audits Forecasting Collaboration, & Uhlmann, E.L. (2022). *On the trajectory of discrimination: A meta-analysis and forecasting survey capturing 44 years of field experiments on gender and hiring decisions*. Manuscript under review.

Scheel, A. M., Schijen, M. R., & Lakens, D. (2021). An excess of positive results: Comparing the standard psychology literature with Registered Reports. *Advances in Methods and Practices in Psychological Science*, *4*(2), 25152459211007467.

Schimmack, U. (2019). Anti-Black bias on the IAT predicts pro-Black bias in behavior. Retrieved March 25, 2022 from https://replicationindex.com/2019/11/24/iat-behavior/

Schooler, J. W. (2011). Unpublished results hide the decline effect. *Nature, 470*(7335), 437.

Schweinsberg, M., Feldman, M., Staub, N., van den Akker, O., van Aert, R., van Assen, M., Liu, Y., … & Uhlmann, E. L. (2021). Radical dispersion of effect size estimates when independent scientists operationalize and test the same hypothesis with the same data. *Organizational Behavior and Human Decision Processes, 165,* 228-249.

Shanks, D. R., Malejka, S., & Vadillo, M. A. (2021). The challenge of inferring unconscious mental processes. *Experimental Psychology, 68*(3), 113-129.

Silberzahn, R., Uhlmann, E. L., Martin, D., Anselmi, P., Aust, F., Awtrey, E., et al.,… & Nosek, B.A. (2018). Many analysts, one dataset: Making transparent how variations in analytical choices affect results. *Advances in Methods and Practices in Psychological Science*, *1*, 337–356.

Simons, D. J. (2014). The value of direct replication. *Perspectives on Psychological Science, 9*(1), 76-80.

Simonsohn, U. (2013). Just post it: the lesson from two cases of fabricated data detected by statistics alone. *Psychological Science*, *24*, 1875–1888.

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2020). Specification curve analysis. *Nature Human Behaviour*, *4*, 1208–1214.

Smith, C. T., & Nosek, B. A. (2011). Affective focus increases the concordance between implicit and explicit attitudes. *Social Psychology, 42*, 300-313.

Smith, C. T., Ratliff, K. A., & Nosek, B. A. (2012). Rapid assimilation: Automatically integrating new information with existing beliefs. *Social Cognition, 30*, 199–219.

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science, 11*(5), 702–712.

Strack, F., Schwarz, N., Bless, H., Kübler, A., & Wänke, M. (1993). Awareness of the influence as a determinant of assimilation versus contrast. *European Journal of Social Psychology, 23,* 53–62.

Tetlock, P. E., Mellers, B. A., Rohrbaugh, N., & Chen, E. (2014). Forecasting tournaments: Tools for increasing transparency and improving the quality of debate. *Current Directions in Psychological Science*, *23*(4), 290-295.

Tierney, W., Hardy, J. H., III., Ebersole, C., Leavitt, K., Viganola, D., Clemente, E., Gordon, M., Dreber, A.A., Johannesson, M., Pfeiffer, T., Hiring Decisions Forecasting Collaboration, & Uhlmann, E. (2020). Creative destruction in science. *Organizational Behavior and Human Decision Processes, 161,* 291-309.

Tierney, W., Cyrus-Lai, W., et al.,… & Uhlmann, E.L. (2022). *Who respects an angry woman? A pre-registered re-examination of the relationships between gender, emotion expression, and status conferral*. Unpublished manuscript.

Uhlmann, E.L. (2014). The problem of the null in the verification of unconscious cognition. *Behavioral and Brain Sciences, 37,* 42-43.

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.

Uhlmann, E.L., & Cohen, G.L. (2007). "I think it, therefore it's true": Effects of self perceived objectivity on hiring discrimination. *Organizational Behavior and Human Decision Processes, 104,* 207-223.

Uhlmann, E.L., & Cohen, G.L. (2005). Constructed criteria: Redefining merit to justify discrimination. *Psychological Science*, *16*, 474-480.

Uhlmann, E.L., Leavitt, K., Menges, J.I., Koopman, J., Howe, M.D., & Johnson, R.E. (2012). Getting explicit about the implicit: A taxonomy of implicit measures and guide for their use in organizational research. *Organizational Research Methods, 15,* 553-601.

Uhlmann, E.L., Pizarro, D.A., & Bloom, P. (2008). Varieties of social cognition. *Journal for the Theory of Social Behaviour, 38*, 293-322.

Varnum, M. E. & Grossmann, I. (2017). Cultural change: The how and the why. *Perspectives on Psychological Science*, *12*(6), 956-972.

Vohs, K.D., Schmeichel, B.J., Lohmann, S., Gronau, Q.F., Finley, Anna J., Wagenmakers, E-

J., et al… & Albarracín, D. (2021). A multi-site preregistered paradigmatic test of the ego depletion effect. *Psychological Science, 32*(10), 1566–1581.

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H. L., & Kievit, R. A. 2012. An agenda for purely confirmatory research. *Perspectives on Psychological Science, 7*(6), 632-638.