Check for updates

Perspectives on Psychological Science
2020, Vol. 15(6) 1329–1345
© The Author(s) 2020
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/1745691620931492
www.psychologicalscience.org/PPS

The Bias of Individuals (in Crowds): Why Implicit Bias Is Probably a Noisily Measured Individual-Level Construct

Paul Connor[1] and **Ellen R. K. Evers**[2]
[1]Department of Psychology, University of California, Berkeley, and [2]Haas School of Business, University of California, Berkeley

Abstract

Payne, Vuletich, and Lundberg's bias-of-crowds model proposes that a number of empirical puzzles can be resolved by conceptualizing implicit bias as a feature of situations rather than a feature of individuals. In the present article we argue against this model and propose that, given the existing evidence, implicit bias is best understood as an individual-level construct measured with substantial error. First, using real and simulated data, we show how each of Payne and colleagues' proposed puzzles can be explained as being the result of measurement error and its reduction via aggregation. Second, we discuss why the authors' counterarguments against this explanation have been unconvincing. Finally, we test a hypothesis derived from the bias-of-crowds model about the effect of an individually targeted "implicit-bias-based expulsion program" within universities and show the model to lack empirical support. We conclude by considering the implications of conceptualizing implicit bias as a noisily measured individual-level construct for ongoing implicit-bias research. All data and code are available at https://osf.io/tj8u6/.

Keywords

implicit bias, social cognition, intergroup relations, measurement, context effects

Few contemporary psychological theories have been as influential as implicit bias. Since its early demonstrations (Fazio, Jackson, Dunton, & Williams, 1995; Greenwald, McGhee, & Schwartz, 1998), it has been the focus of an enormous body of research (Amodio & Mendoza, 2010; Greenwald, Poehlman, Uhlmann, & Banaji, 2009; Oswald, Mitchell, Blanton, Jaccard, & Tetlock, 2013), measured in millions of online volunteers (Xu, Nosek, & Greenwald, 2014), and discussed during presidential debates (Weir, 2016). However, various issues have left implicit-bias researchers divided on how to conceptualize the construct. In their 2017 article, Payne, Vuletich, and Lundberg (2017a) introduced a new way of thinking about the phenomenon via their bias-of-crowds model. In contrast to previous conceptualizations that assume implicit bias to represent a relatively stable measure of an individual-level variable, the authors argued that implicit bias should primarily be understood as a stable feature of situations rather than of persons. By reconceptualizing implicit bias in this way, the authors proposed that a number of empirical puzzles could be solved.

The present article argues against the bias-of-crowds model. First, we outline Payne and colleagues' theory, as well as supporting evidence, and the empirical puzzles it was proposed to solve. Second, we show using both real and simulated data how these empirical puzzles can be parsimoniously explained as the expected outcomes of individual-level measurement error and its reduction via aggregation. Third, we discuss why Payne and colleagues' counterarguments for dismissing measurement error as an explanation for these puzzles have been unconvincing. Finally, we draw on publicly available implicit-bias data to investigate a testable implication of the model and find it to lack empirical support. Given the available evidence, we conclude that implicit bias remains best conceptualized as a noisily measured individual-level construct, albeit one that, like most individual-level psychological constructs, can be affected

Corresponding Author:
Paul Connor, 2121 Berkeley Way, University of California, Berkeley, CA 94720
E-mail: pconnor@berkeley.edu

by features of situations. We then consider the implications of this conceptualization for implicit-bias research.

The Bias-of-Crowds Model, the Puzzles, and the Evidence

The bias-of-crowds model (Payne et al., 2017a) proposes that implicit bias should be understood as being primarily an aspect of social situations rather than of individuals. To support this contention, Payne and colleagues presented three puzzles that can best be explained, they argued, by adopting this view.

First, they observed that implicit bias appears to be "large and unstable" (Payne et al., 2017a, p. 233), meaning that although implicit bias is robustly replicable (e.g., sample means reliably demonstrate greater population-level automatic associations between "Black" and "bad" and between "White" and "good"), it is also relatively unstable at the individual level; scores on repeated tests correlate weakly (the meta-analytic test–retest reliability figure the authors provide is r = .42, from Gawronski, Morrison, Phills, & Galdi, 2017).

Second, the authors describe how implicit bias appears to be "permanent yet unstable" (Payne et al., 2017a, p. 234), meaning that despite the volatility of individuals' scores from test to test, mean levels of implicit bias have been observed to be relatively similar in individuals of all ages, from children to older adults (Baron & Banaji, 2006).

Third, Payne and colleagues describe the puzzle of "places and people" (Payne et al., 2017a, p. 234), referring to the observation that implicit-bias scores correlate weakly with related constructs (e.g., discriminatory behaviors) at the individual level but appear to correlate more strongly with these constructs at the level of geographic regions. For example, meta-analyses have estimated the individual-level correlation (r) between implicit-bias scores and discriminatory behaviors to be as low as .14 (Oswald et al., 2013), but a number of studies have reported much higher correlations between implicit bias and indicators of discriminatory behaviors at the group level—for example, between U.S. states' average levels of implicit bias and their relative frequencies of Internet searches involving racial slurs (r = .78, Rae, Newheiser, & Olson, 2015).

These, then, are the puzzles that the bias-of-crowds model was proposed to resolve: (a) Implicit bias is unstable yet robustly replicable, (b) implicit bias varies within individuals across measurement occasions but is stable at the group level, and (c) implicit bias correlates relatively weakly with related constructs at the individual level but strongly at the group level.

To solve these puzzles, Payne and colleagues (2017a) began by noting that although most conceptualizations consider implicit bias to be primarily an individual-level construct, there is also a general consensus that implicit bias can be affected by features of situations. For example, implicit-racial-bias scores have been shown to be affected by interacting with a Black experimenter, listening to rap music, or looking at photos of Black celebrities (for review, see Lai, Hoffman, & Nosek, 2013). The bias-of-crowds model departs from past conceptualizations, however, by proposing that implicit bias is primarily a feature of situations rather than of individuals. Thus, just as no implicit-bias researcher would deny that implicit bias can be affected by features of situations, the bias-of-crowds model does not deny that implicit bias is to some extent an attribute of individuals. Payne and colleagues (2017a) wrote,

> To summarize the view put forward here, although implicit bias can in principle exist as an attribute of persons or an attribute of situations, the empirical evidence is more consistent with the situational view. By switching the emphasis from a person-based analysis to a situation-based view, we arrive at a reinterpretation of the empirical data. This new interpretation suggests that measures of implicit bias are meaningful, valid, and reliable. Contrary to most assumptions, however, they are meaningful, valid, and reliable measures of situations rather than persons. (p. 236)

A Parsimonious Alternative: Measurement Error and Aggregation

Should researchers adopt the bias-of-crowds model and reconceptualize implicit bias as primarily a feature of situations rather than of individuals? We can certainly understand the impulse to do so. Since its inception, two of the most sustained and damaging critiques of implicit-bias research have been the psychometric unreliability of its measures and its relatively weak correlation with associated constructs (e.g., Oswald et al., 2013; Singal, 2017). The bias-of-crowds model offers a seemingly powerful response to such criticisms, painting implicit bias as "meaningful, valid, and reliable" (Payne et al., 2017a, p. 236)—but in the context of situations, not of persons.

Yet despite how appealing the model may be, it is important to consider whether such a reinterpretation is supported by the evidence. And in doing so, it is important to acknowledge that a parsimonious alternative explanation is readily available for each of the puzzles described by Payne and colleagues (2017a). Instead of reconceptualizing implicit bias as a feature of situations, this alternate view simply requires conceptualizing implicit bias as being an individual-level construct measured with substantial measurement error. In this view, although implicit bias is a relatively stable feature of individuals, scores on implicit-bias tests do not accurately reflect individuals' stable levels of bias. Thus, it is natural we would see low test–retest and criterion correlations for implicit-bias scores as a result of the error associated with individual measurements (Furr & Bacharach, 2013). In addition—although this is perhaps less obvious—it is also natural that we would see high aggregate-level correlations. When enough noisy individual-level scores are aggregated, positive and negative measurement errors tend to cancel each other out, resulting in highly accurate measures of group means. Assuming that some real differences exist among group means (i.e., relatively more or less biased individuals clustering together in specific groups), this heightened measurement accuracy at the group level will tend to produce exactly the observations described by Payne and colleagues' puzzles: higher stability across measurement occasions for group means than for individual scores (the "large and unstable" and "permanent yet unstable" puzzles) and higher correlations between related constructs at the group level compared with the individual level (the "places and people" puzzle).

These points were touched on in two of the immediate responses offered to the bias-of-crowds model (Kurdi & Banaji, 2017; Rae & Greenwald, 2017). However, we believe these responses failed to fully elucidate the extent to which measurement error undermines the key motivations behind the bias-of-crowds model. The model was introduced primarily to explain greater stability among group means than individual scores, and greater correlations with related constructs at the group level than at the individual level. These were the observations that led Payne and colleagues to conclude that "most of the systematic variance in implicit bias is situational" (2017a, p. 233). However, as we will show, if we fully consider the role of measurement error and the way it is reduced by aggregation, we can see that (a) even extremely strong correlations at the group level can represent relatively trivial amounts of systematic variance compared with individual differences and (b) each of these puzzles is completely compatible with a parsimonious alternative account conceptualizing implicit bias as an individual-level construct measured with substantial error.

The bias of weekdays: why seemingly large aggregate-level effects can be misleading

To demonstrate why higher correlations at the group level can be misleading, we used data from the Project Implicit race IAT data set (Xu et al., 2014), which contains measurements of the implicit racial bias of 3,432,939 U.S.-based volunteer participants measured on the race IAT (Greenwald et al., 1998; Greenwald, Nosek, & Banaji, 2003) between 2004 and 2018.[1] Just as Payne and colleagues (2017a) note, Project Implicit's data suggest that race-IAT scores are unstable at the individual level. Among 1,213 identifiable U.S. cases measured on multiple occasions,[2] the test–retest correlation (*r*) is .24 (although this is likely an underestimate; we discuss various estimates of the race IAT's test–retest reliability below in discussing our simulation results). The observed relationship between implicit bias and related criteria at the individual level is also weak: IAT scores correlate at .30 with participants' "thermometer bias" scores (explicit ratings of warmth toward Whites minus ratings of warmth toward Blacks). However, just as Payne and colleagues observed, aggregating IAT scores within states produces stronger correlations. The test–retest correlation between U.S. states' average IAT scores for 2017 and 2018 is .89, and the overall state-level correlation between IAT scores and thermometer bias is .55.

Yet although these state-level correlations may seem impressively large, they do not represent "most of the systematic variance" in implicit bias. In fact, only a small amount of variance in implicit bias is attributable to variation between states. In the Project Implicit data, a linear model predicting IAT scores from fixed effects of participants' states achieves an r^2 of .0026. This means that participants' states of residence account for just 0.26% of the total variation in IAT scores. By contrast, on the basis of the observed individual-level test–retest correlation of .24 reported above, participants' previous observed scores explain around 6% of the total variation in IAT scores (if r = .24, r^2 = .0576), and—assuming momentarily that the test–retest reliability 0.24 is accurate—participants' true stable levels of bias, if observable, would explain 24% of total variation in measured scores (test–retest reliability represents the squared correlation between true and measured scores; see Cohen, Cohen, West, & Aiken, 2003). Between states and individuals, then, most of the systematic variance of IAT scores within the Project Implicit data resides at the individual level, not the state level.

Why, then, have Payne and colleagues claimed that most of the systematic variance in implicit bias is at the level of the situation rather than the individual? Another possible interpretation is that this phrase was not meant to refer to total amounts of variance explained, but to amounts of variance explained at specific levels of aggregation. This argument could be formalized as follows: Regardless of the total amount of variation explained at each level of analysis, the level of analysis at which variables exhibit the highest reliability and criterion correlations is the level at which most of the systematic variance lies.

However, this position produces some strange conclusions. For example, the Project Implicit data also

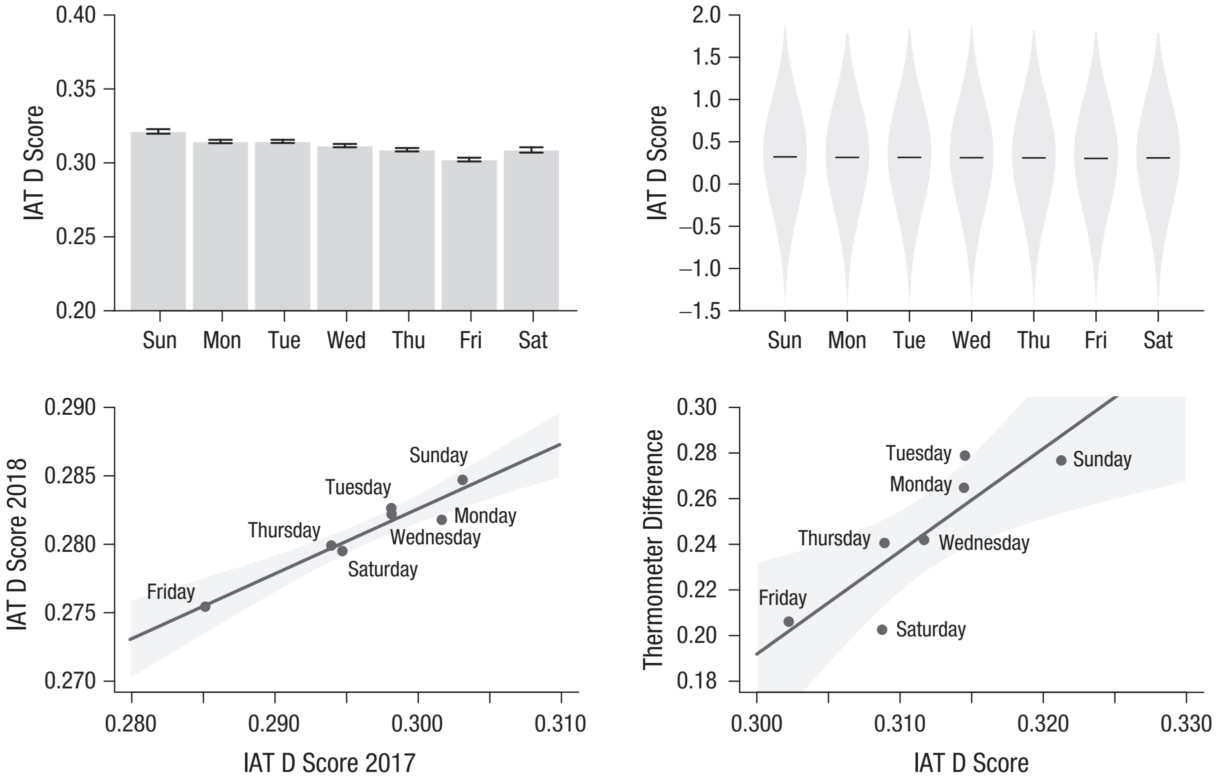

Fig. 1. Results of aggregating implicit bias data at the weekday level. The top panels show the differences in Project Implicit Implicit Association Test (IAT) D scores (see Note 4) by weekday (top left) and in relation to the overall distribution of IAT D scores (top right). The bottom panels show the relationship between average weekday IAT D scores from 2017 and 2018 (bottom left) and between thermometer bias and average weekday IAT D scores (bottom right). Bars in top panels and shaded regions in bottom panels indicate 95% confidence intervals.

show that implicit bias appears to vary according to the day of the week. The relationship is robustly significant, $F(6, 3,432,932) = 75.36$, $p < .001$, but is also extremely weak, with weekdays explaining just 0.01% of variation in IAT scores. This suggests a modest conclusion: Weekdays may have some relationship with implicit bias, but the relationship, if it exists, is relatively trivial. Yet if we aggregate IAT scores within weekdays, we can tell a very different story. Aggregated IAT scores for weekdays from 2017 and from 2018 correlate nearly perfectly ($r = .95$; see Fig. 1, bottom left panel), and despite the low individual-level correlation in the data between implicit and thermometer bias, aggregating scores for both variables within weekdays produces a correlation of 86 (see Fig. 1, bottom right panel).

But does this mean that most of the systematic variance in implicit bias is in fact at the weekday level, and so implicit bias should be reconceptualized as being primarily a feature of weekdays? If we follow the logic of the bias-of-crowds model, this seems the unavoidable conclusion. Despite the fact that weekdays account for just 0.01% of variation in IAT scores, and individuals' previous scores account for around 6%—600 times more—aggregating scores within weekdays produces both greater test–retest correlations and criterion correlations at the weekday level than at the individual level. So if what matters for how we conceptualize implicit bias is the level at which we can observe the greatest reliability and criterion correlations, then we must conclude that most of the systematic variance in implicit bias is at the weekday level.

We hope that readers agree that this would be a strange conclusion. Weekdays may vary in terms of average implicit bias, just as states do. And by aggregating large numbers of noisy individual-level scores within weekdays, we may be able to measure these differences extremely accurately, just as we can with states. But we should not be so impressed by high weekday-level correlations that we deem them to contain most of the systematic variance in implicit bias. If enough scores are aggregated, weekday-level implicit bias from 2017 can explain a large proportion of the weekday-level variation in implicit bias in 2018. But this predictive power must be interpreted in recognition of the fact that there is very little weekday-level variation in implicit bias to be explained. Thus, although aggregating large numbers of scores may produce apparently larger effects when expressed as correlations, such effects can in fact be relatively trivial in the degree to which they explain total observed variance.

Are the "puzzles" puzzling?

To provide a more formal demonstration of how measurement error and aggregation can combine to explain each of the puzzles discussed by Payne and colleagues (2017a), we used computer simulation. Using the R software environment (Version 3.6.1; R Core Team, 2019), we simulated data sets in which the alternate view we have described above—that implicit bias is a noisily measured individual-level construct—was known to be true. Namely, we specified in simulated data sets that (a) implicit bias is a stable feature of individuals; (b) implicit bias is correlated with other criteria at the individual level; (c) there are group-mean differences in implicit bias, and some groups contain relatively more or fewer biased individuals; and (d) implicit bias is measured with error. With each of these conditions met, we examined the extent to which each of the empirical puzzles discussed by Payne and colleagues (e.g., higher test–retest and criterion correlations at the group level than at the individual level) occurred within the simulated data.

The technical details of our simulations were as follows. We began by simulating normally distributed vectors of individual-level true scores of implicit bias (we refer to these vectors as *true*), and assigning each score to one of 10 groups of equal size. At this point we set two key parameters: per-group *N*, which determined the number of cases in each group, and the intraclass correlation (ICC) of the true scores, which represents the proportion of the total variance of the true scores explained by group membership (higher ICCs indicate more variation between groups relative to variation within groups). Following this, we simulated noisily measured scores of implicit bias at two different time points, test and retest. We then specified a third parameter, $r_{\text{true,measured}}$, which determined the individual-level measurement error by setting the correlation between individuals' true levels of implicit bias (the true scores) and each of the measured implicit-bias scores (test and retest). Finally, we simulated normally distributed scores on a related criterion (which we call *criterion*) and set a final parameter, $r_{\text{true,criterion}}$, which determined the correlation between individuals' true levels of implicit bias (the true scores) and the criterion scores.

Our simulation process therefore allowed us to systematically vary four key parameters: (a) the size of groups being aggregated (per-group *N*), (b) the ICC of the true implicit-bias scores, (c) the individual-level correlation between true and measured implicit-bias scores ($r_{\text{true,measured}}$), and (d) the individual-level correlation between true implicit-bias scores and scores on a criterion ($r_{\text{true,criterion}}$). We selected ranges of values for each parameter: for per-group *N*, 10, 50, 250, and 1,000; for ICC, 0.001, 0.01, 0.05, and 0.1; for $r_{\text{true,measured}}$, .1, .3, .5, .7, and .9; and for $r_{\text{true,criterion}}$, .1, .3, .5, .7, and .9. Then, across 100 iterations, we simulated data sets consisting of true, test, retest, and criterion scores for each unique combinations of parameters.

For each of the simulated data sets, we computed four key outcomes of interest: (a) individual-level test–retest reliability, defined as the correlation between test and retest; (b) aggregate-level test–retest reliability, defined as the group-level correlation between test and retest; (c) individual-level criterion correlation, defined as the correlation between test and criterion; and (d) aggregate-level criterion correlation, defined as the group-level correlation between test and criterion.[3]

An example of a single data set created via this simulation process is depicted in Figure 2. In this simulated data set, the per-group *N* is 500, the correlation (*r*) between measured and true scores is set to .5, and the true ICC is set to .05. In these data, then, just 5% of variation in true implicit-bias scores is attributable to differences between groups; the other 95% represents individual differences. Nonetheless, because of the measurement error associated with the measured scores, there is a much higher test–retest correlation at the aggregate level ($r = .85$) than at the individual level ($r = .24$).

Test–retest reliability. Figure 3 displays the test–retest reliabilities at the individual and aggregate level obtained for different levels of measurement error ($r_{\text{true,measured}}$), ICC, and per-group *N*. The red lines show that, unsurprisingly, individual-level test–retest reliability was purely a function of measurement error. The blue lines, however, show that aggregate-level test–retest reliability depended not only on measurement error but also on the ICC and per-group *N*. When ICC was low (i.e., the top row of plots), or sample sizes were low (i.e., the leftmost column of plots), there was little difference between test–retest reliabilities at the individual and aggregate levels. However, as ICCs increased, aggregation of larger groups created much higher test–retest reliabilities at the aggregate level than at the individual level.

Criterion correlations. Similar results obtained for criterion correlations. Figure 4 displays criterion correlations at the individual and aggregate levels for different levels of measurement error ($r_{\text{true,measured}}$), ICC, per-group *N*, and correlation between individuals' true levels of implicit bias and the criterion ($r_{\text{true,criterion}}$). The various

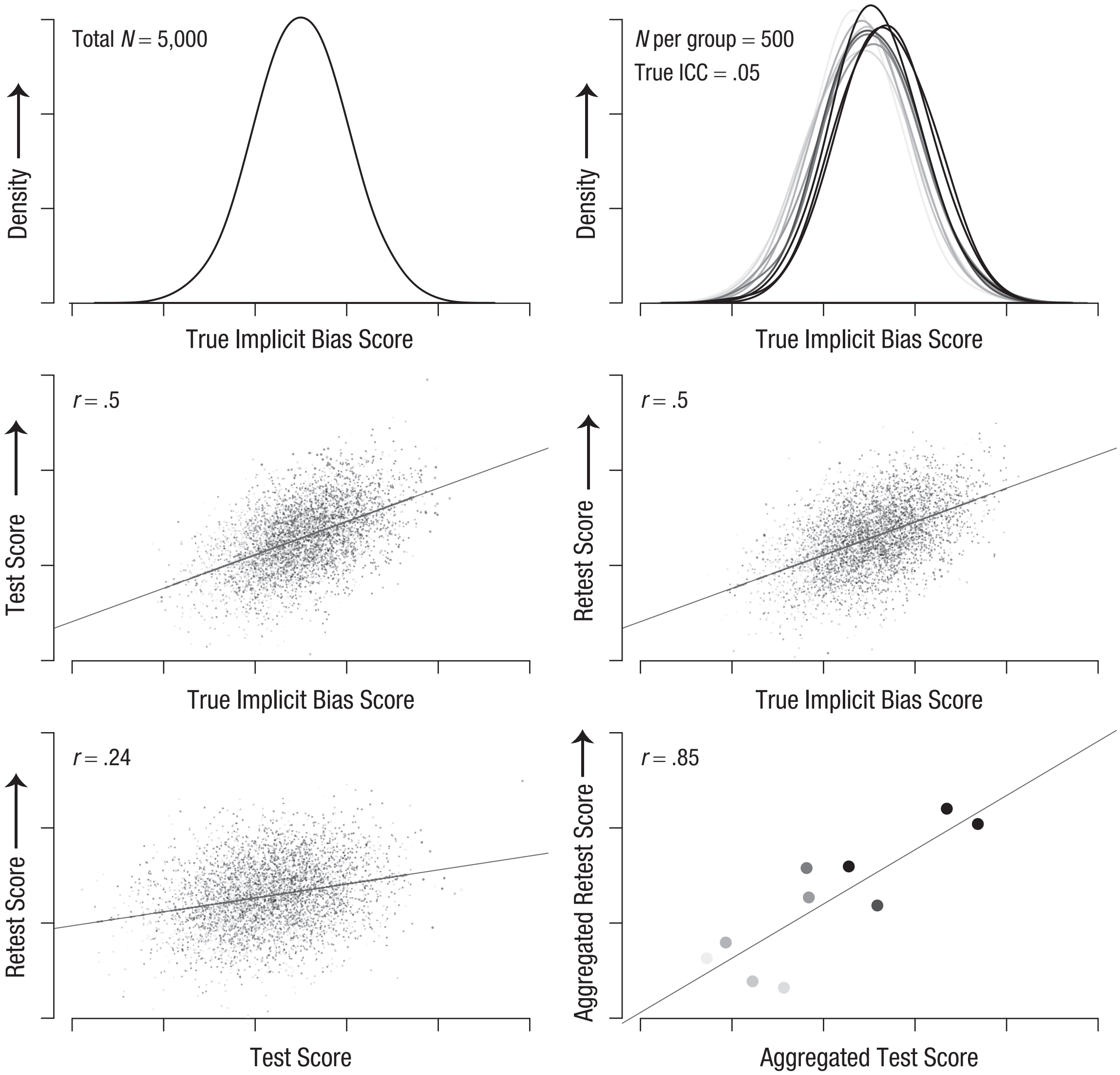

Fig. 2. One simulated data set. The top panels display the simulated distribution of true implicit-bias scores overall (left) and by group (right). The middle panels display scatterplots (with best-fitting regression lines) for the relationships between true and measured scores for test (left) and retest (right). The bottom panels display scatterplots (with best-fitting regression lines) for the resulting test–retest correlations at the individual level (left) and group level (right).

shades of red lines show that individual-level criterion correlations were a function of both (a) the strength of the true criterion correlation and (b) measurement error. The various shades of blue lines, however, show that aggregate-level criterion correlations depended not only on the strength of the true criterion correlations and the measurement error but also on ICC and per-group *N*. When ICC was low or per-group *N* was low, there was little difference between criterion correlations at the individual and aggregate levels. However, as ICCs increased, aggregation of larger groups created substantially higher criterion correlations at the aggregate level than at the individual level.

Implications for race-IAT scores. To assess what these results may mean for the specific case of implicit racial bias, we need to ask, "what are the relevant values of the intraclass correlation and measurement accuracy?" Observed ICCs of implicit racial bias have varied, but they are typically low. For example, we reported above that around 0.25% of variance in race-IAT scores in the Project Implicit data can be explained by variation between states, whereas Vuletich and Payne (2019) reported 1% of variance within data gathered by Lai and colleagues (2016) to be attributable to differences between universities. Measurement error also reduces observed ICCs below their true levels—the ICCs plotted above in Figures 3 and

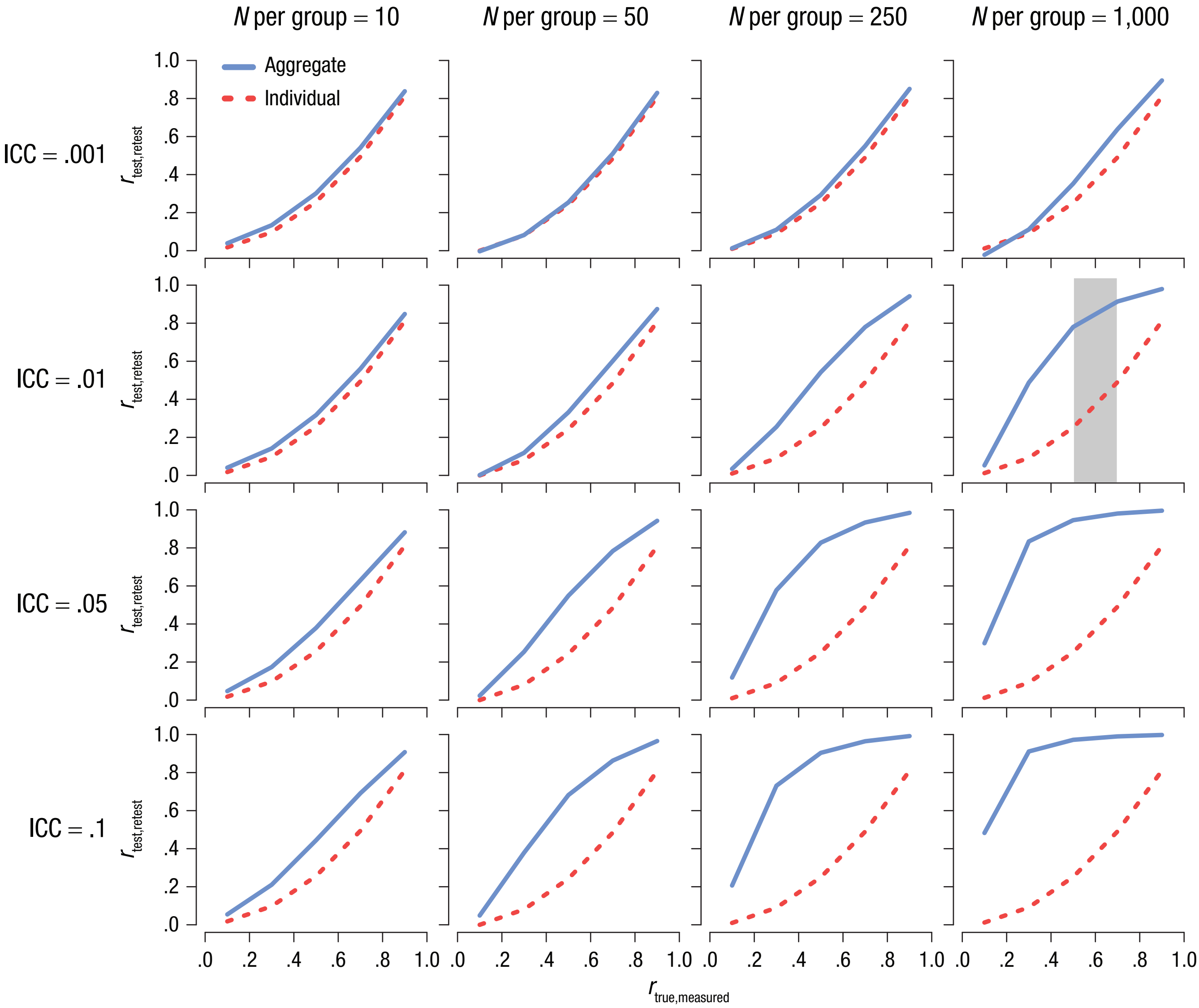

Fig. 3. Simulated test–retest reliabilities at the individual and aggregate levels at differing levels of measurement accuracy, true intra-class correlation, and sample size per group. For example, the shaded area has a per-group N = 1,000, ICC = .01, and $r_{\text{true,measured}}$ = .5 to .7.

4 refer to the true ICCs, not the measured ICCs. Given these considerations, the most relevant plots to consider in Figures 3 and 4 are likely those in the second row, in which the true ICC is set to .01; this level of true ICC will yield observed ICCs of .01 and lower, depending on the level of measurement error.

Measurement accuracy can be assessed via estimates of test–retest reliability. Recent meta-analytic work puts the test–retest reliability (*r*) of the IAT at .5 (Greenwald & Lai, 2020); however, there is evidence that the race IAT (as opposed to, say, a self-esteem IAT) may be less reliable than this. Gawronski and colleagues (2017) reported 10 test–retest correlations for the race IAT for which the weighted mean was .38. And as discussed earlier, in the Project Implicit data (which we believe is the largest sample to have been measured on the full IAT measure on multiple occasions), we observed a test–retest reliability of .24. This figure is likely to be an underestimate, however, because individuals voluntarily retaking Project Implicit's race IAT may have had a reason to believe there was a problem with their original scores. In support of this, we found Project Implicit scores on retests to be significantly lower (*M* = 0.19, *SD* = 0.44) than scores on original tests (*M* = 0.30, *SD* = 0.46), *t*(1212) = 7.28, *d* = 0.26, suggesting that participants may have (a) had insight that their original score was an overestimate of their bias, (b) deliberately tried to reduce their scores on retests, or

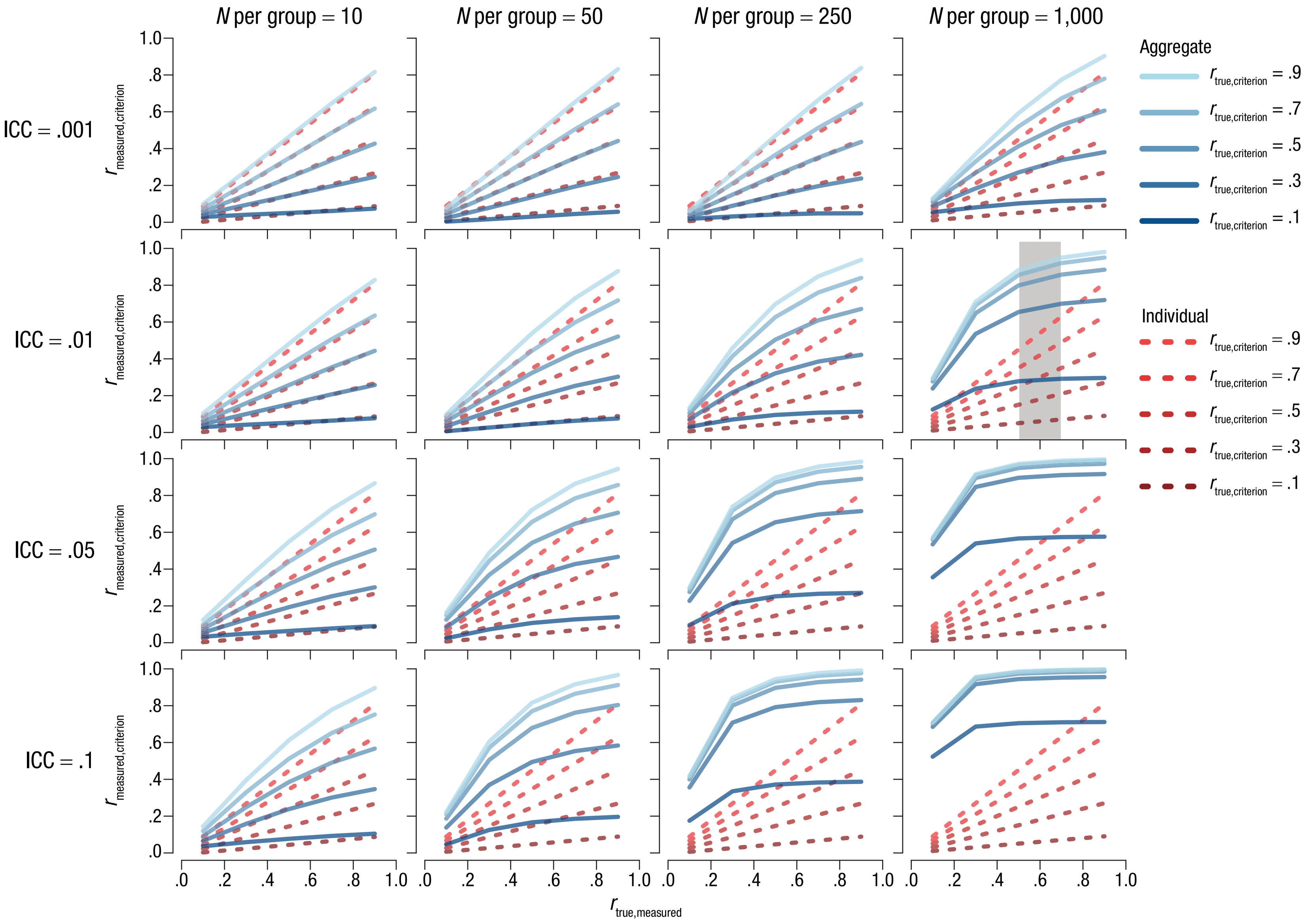

Fig. 4. Simulated criterion correlations at the individual and aggregate levels at differing levels of measurement accuracy, true intra-class correlation, correlation between true scores and criterions, and sample size per group. As in Fig. 3, the shaded area has per-group N = 1,000, ICC = .01, and $r_{\text{true,measured}}$ = .5 to .7.

(c) both. Nonetheless, it seems safe to conclude that the test–retest reliability of the race IAT probably lies somewhere between .25 and .5, which implies a correlation between measurements and individuals' true scores of somewhere between .5 and .71.

By narrowing our focus in this way, we can see that measurement error and aggregation can easily explain each of the empirical puzzles concerning race-IAT scores discussed by Payne and colleagues. As shown in Figure 3, with true ICCs of 0.01 and correlations (*r*) of .5 and .7 between true and measured implicit-bias scores, aggregating 1,000 scores per group produced average individual-level test–retest correlations of .25 and .49, respectively, but average aggregate-level test–retest correlations of .78 and .91, respectively (see the shaded region in Fig. 3). And at the same levels of measurement accuracy, when the true correlation between implicit bias and the criterion was set to .3, aggregating 1,000 scores per group produced average individual-level criterion correlations of .15 and .21, respectively, but average aggregate-level criterion correlations of .65 and .70, respectively (see the shaded region in Fig. 4).

Therefore, given levels of measurement accuracy and ICCs similar to those observed in the case of implicit bias, aggregation of noisy individual-level scores produced both greater stability of group means compared with individual scores (Fig. 3) and greater correlations with related criteria (Fig. 4) at the group level than at the individual level. Payne and colleagues' empirical puzzles are therefore not puzzling at all. Rather, they are exactly what we would expect to see if implicit bias was a noisily measured individual-level construct.

The Counterarguments

Payne and colleagues have offered two main counterarguments as to why measurement error is an insufficient explanation for the puzzles (Payne et al., 2017a; Vuletich & Payne, 2019). In what follows we respond to each of these arguments.

The internal-consistency argument

Payne and colleagues (2017a) have noted that although implicit-bias test scores are unstable over time, they nonetheless exhibit relatively high levels of internal consistency. This internal consistency, they argue, means that although implicit-bias tests capture individuals' biases accurately at each time point, individuals' biases themselves are unstable over time. They write, "given that internal consistency cannot easily explain the low temporal stability of implicit bias measures, the most likely explanation is that the unreliability lies in the malleability of people's psychological biases rather than in the tests" (Payne et al., 2017a, p. 234).

The internal consistency of implicit-bias tests is indeed greater than their test–retest reliability. A recent meta-analytic estimate put the overall internal consistency of the IAT at $r = .80$ (Greenwald & Lai, 2020), and although some evidence suggests—as discussed above—that the psychometric properties of the race IAT may not be quite as strong as those of other variants of the IAT, there is little doubt that the tests exhibit higher internal consistency than test–retest reliability; the meta-analysis authors concluded that test–retest reliability "is generally smaller than [internal consistency]; . . . this difference indicates the presence of systematic variance in [internal consistency] that is unshared among measurement occasions" (p. 436).

Therefore, there does appear to be an important amount of variation in IAT scores that is attributable neither to individuals' chronic, stable levels of bias nor to random measurement error. However, this does not imply that implicit bias is not an individual-level construct. It is normal for individual-level variables to vary across time within individuals. For example, scores on measures of life satisfaction are known to be affected by transient moods and even by sports results (Schwarz, Strack, Kommer, & Wagner, 1987). But this does not mean that life satisfaction is not an individual-level construct. It simply means that individuals' measured scores on the construct at any given time are affected by factors other than individuals' long-term, chronic levels of the construct.

In the case of implicit bias, the causal factors that underlie intraindividual variation over and above measurement error remain largely an empirical question. According to the bias-of-crowds model, intraindividual variation is largely caused by individuals' incidental exposure to structural inequalities within social environments. Such exposure is theorized to make intergroup stereotypes temporarily accessible, and thereby produce implicit bias. Payne and colleagues (2017a) wrote,

> Town A has relatively high levels of systemic racism. Housing patterns and schools are highly segregated, and they are correlated with large disparities in incomes. . . . When police pull over motorists, or criminal suspects are described in the local news, they are disproportionately Black or Hispanic. Town B, in contrast, has low levels of systemic racism. Residents know all the same stereotypes as everyone else. But strolling through the town, residents are unlikely to see those stereotypes confirmed by inequalities in living conditions and social roles on a daily basis. Because of the difference in daily reminders of inequality,

> the average accessibility of stereotypical links will be different in the two towns. Implicit bias will be higher in Town A than Town B. (p. 239)

In support of the theorized causal link between such "daily reminders of inequality" and implicit bias, Payne and colleagues have highlighted associations between mean levels of implicit bias within counties and (a) the prevalence of slavery in counties in 1860, (b) counties' racial disparities in poverty rates, and (c) counties' racial disparities in intergenerational mobility (Payne, Vuletich, & Brown-Iannuzzi, 2019). Further, they have highlighted associations between mean levels of implicit bias within universities and (a) the presence of Confederate statues within universities, (b) racial diversity among universities' faculty, and (c) levels of socioeconomic mobility among students (Vuletich & Payne, 2019).

However, even if we conceptualize implicit bias as an individual-level phenomenon, it is unsurprising that these cross-sectional associations exist, and completely plausible that the causal arrow could be going in precisely the opposite direction—that is, individuals' implicit biases produce structural inequalities. It is easy to see why universities whose staff and students have relatively high levels of implicit bias might be more likely to preserve Confederate statues or hire less diverse faculties and why counties whose residents have relatively high levels of implicit bias might adopt policies that create greater racial segregation or economic disparities.

Moreover, even if we accept for the sake of argument the very plausible hypothesis that there is a causal link from exposure to structural inequalities to individual-level implicit bias, this still does not establish that exposure to structural inequalities is an important cause of intraindividual variation in implicit bias between measurement occasions. To establish this, Payne and colleagues would need to show not only that the structural features of Town A produce relatively higher bias among its residents but also that intraindividual variation in exposure to housing segregation, detained motorists, or racist news stories within Town A is systematically related to intraindividual variation in implicit bias. To our knowledge, there is no such evidence. However, much current evidence suggests that many of the factors likely to underlie intraindividual variation in implicit bias are variables that have little to do with such structural factors. For example, other intraindividual factors known to affect IAT scores include emotional states (Dasgupta, DeSteno, Williams, & Hunsinger, 2009), fatigue (Johnson et al., 2016), and motivation (Devine, Plant, Amodio, Harmon-Jones, & Vance, 2002). It is therefore plausible that factors such as these explain much more of the intraindividual variability in IAT scores than do structural features of environments, and there is currently just as much, if not more, causal evidence for these possibilities compared with Payne and colleagues' speculative interpretations of correlational findings. The mere fact of greater internal consistency than test–retest reliability is therefore completely compatible with the view that implicit bias is an individual-level construct.

The permutation argument

In a follow up article, Vuletich and Payne (2019) reanalyzed the longitudinal data gathered by Lai and colleagues (2016), who obtained multiple race IAT D score[4] measurements, separated by 1 to 4 days, from approximately 5,000 participants across 18 different universities. Using multilevel modeling, the authors predicted IAT scores at Time 2 from IAT scores at Time 1, but decomposed variation in Time 1 scores into within-groups and between-groups variation. Their model was therefore

$$\begin{aligned}\mathrm{IAT}\_t2_{ij} &= \beta_0 + \beta_1\mathrm{IAT}\_t1_j + \beta_2(\mathrm{IAT}\_t1_{ij} - \mathrm{IAT}\_t1_j)\\ &\quad + \eta_j + \varepsilon_{ij},\end{aligned}$$

where i indexes individuals and j indexes universities, IAT_t2_{ij} is the Time 2 IAT score for individual i in university j, IAT_t1_j is the Time 1 mean IAT score for university j, $\mathrm{IAT}\_t1_{ij} - \mathrm{IAT}\_t1_j$ is the deviation of individual i in university j from the university mean, η_j is a random intercept adjustment for university j, and ε_{ij} is the leftover residual.

In this model, then, β_1 is the expected change in individuals' implicit Time 2 bias associated with a 1-unit increase in their university's mean implicit bias at Time 1 ($\beta_1\mathrm{IAT}\_t1_j$), whereas β_2 is the expected change in individuals' Time 2 implicit bias associated with a 1-unit increase in their Time 1 deviation from their university's mean. Observing a larger β_1 slope (0.87) than β_2 slope (0.25), Vuletich and Payne (2019) concluded that the effect of the between-universities variation was "much larger" (p. 5) than the effect of within-universities variation, and interpreted this as support for the bias-of-crowds model:

Vuletich and Payne (2019) considered but rejected the idea that this result was a mere artifact of aggregation. To do so, they used a permutation exercise consisting of the following:

1. Randomly reshuffling the university affiliations of each individual case;
2. redecomposing Time 1 IAT scores into variation between and within the reshuffled universities;

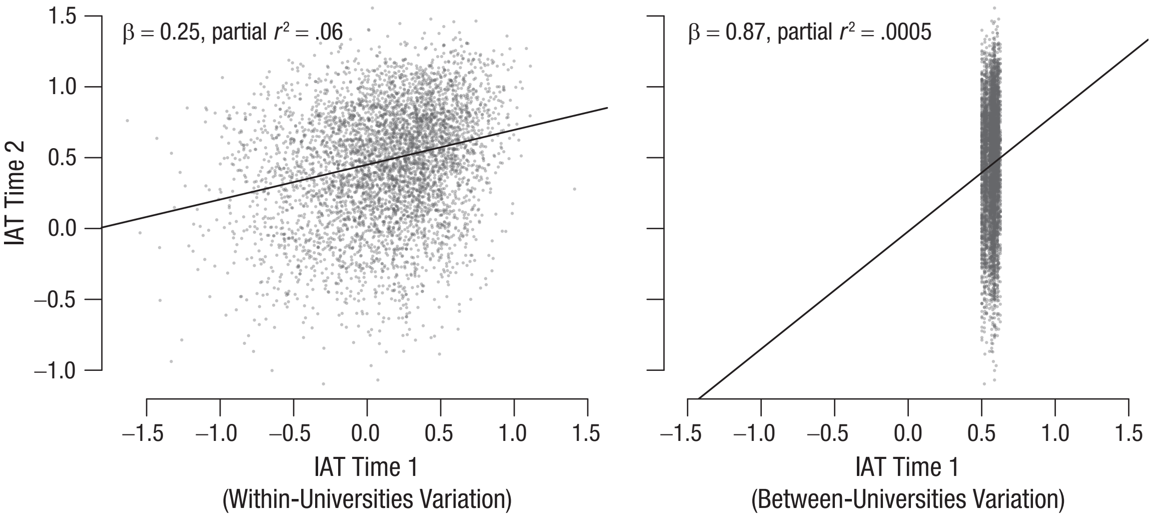

Fig. 5. The relationships between Time 1 and Time 2 IAT scores in Lai and colleagues' (2016) data. Time 1 scores are decomposed into variation within universities (left) and between universities (right).

3. again predicting Time 2 IAT scores from the recalculated between- and within-universities variation via the above model, recording the resulting β_1 and β_2 slopes; and
4. repeating Steps 1 to 3 one hundred times.

They observed the average between-universities slope to be reduced compared with their original model, and similar in magnitude to the original within-universities slope ($\overline{\beta_1}$ = 0.26). On the basis of this result, they argued that their original result could not have been due to aggregation alone:

> Whereas our original analysis revealed a large campus-level effect and a small person-level effect, randomly assigning the persons to nominal groups left only systematic effects at the person level. The large reduction of the campus-level effect suggests that there was indeed a campus-specific signal that was revealed by aggregation. (p. 5)

However, there are numerous problems with these arguments. First, in this context, a larger β_1 than β_2 slope does not indicate a larger effect in terms of total variation explained, but merely a greater aggregate-level correlation than individual-level correlation. And as we showed above, correlations at alternate levels of analysis can be misleading when there is much more variation at one level than another (recall the weekday example). This is the case in Lai and colleagues' (2016) data: There is vastly more variation within universities than between universities. The within-universities standard deviation—i.e., the standard deviation of the predictor (IAT_t1_{ij} – IAT_t1_j)—is 0.41, whereas the between-universities standard deviation—i.e., the standard deviation of the predictor IAT_t1_j—is just 0.03. Thus, although the β_1 slope predicts a very large change in individuals' Time 2 scores for a 1-unit change in between-universities variation, this slope must be interpreted with care, because university mean IAT scores do not differ from each other by 1 unit. The two most different universities in terms of Time 1 means were the University of California, Irvine (M = 0.50), and University of Texas at Austin (M = 0.63), whose means differed by just 0.13 units. By contrast, individuals within universities *do* routinely differ from each other by 1 unit. Of all students measured at Time 1, 99% of cases differed by at least 1 unit in IAT score from another student at their own university.

This is why we get a very different answer regarding which effect is larger if we examine variance explained rather than slopes. We repeated Vuletich and Payne's analysis and reproduced their exact between- and within-universities slopes. However, when we examined partial r^2 values for each effect (Nakagawa & Schielzeth, 2013), we found the within-groups effect (partial r^2 = .058) to explain around 12 times as much variance in Time 2 scores as the between-groups effect (partial r^2 = .005), despite its smaller slope. Figure 5 displays this discrepancy. The between-groups slope (Fig. 5, right) is steeper, but because of the lack of variation between groups, it explains far less variation than the within-groups effect (Fig. 5, left).

To provide another demonstration that such results can stem directly from measurement error and aggregation, we analyzed Lai and colleagues' (2016) measures of explicit racial bias, which was also measured at

Times 1 and 2. These explicit-bias scores demonstrated a high test–retest reliability (r = .86), so we assume it is uncontroversial to consider them measures of an individual-level construct. To test whether Vuletich and Payne's results would succeed for a noisily measured individual construct, we artificially added measurement error to these explicit-bias scores by iteratively adding normally distributed random noise (M = 0) to the raw scores until their test–retest correlation fell below that of the implicit-bias scores (r < .246). We then repeated Vuletich and Payne's (2019) variance decomposition and modeling approach on these artificially noisy explicit-bias scores. By repeating this process 1,000 times, we found the average between-groups slope to be both substantially larger ($\bar{\beta}_1$ = 0.65) and more variable (SD = 0.17) than the average within-groups slope ($\bar{\beta}_2$ = 0.24, SD = 0.004). Thus, the greater slope on the between-groups variation than on the within-groups variation observed by Vuletich and Payne is completely compatible with implicit-bias scores representing an individual-level variable measured with substantial error.

Vuletich and Payne's (2019) permutation result is also compatible with this alternative explanation. Group-level test–retest correlations rely on the existence of systematic group-level variation, meaning that relatively more or less biased individuals must be clustered together, at least to some extent, within specific groups. Shuffling individuals between groups removes systematic group-level variation, and thus naturally reduces group-level test–retest correlations. To demonstrate this, we again added random error to Lai and colleagues' (2016) explicit bias measures and, by trial and error, found a configuration giving us a test–retest reliability (r = .24), between-groups slope (β_1 = 0.86), within-groups slope (β_1 = 0.24), and ICC (0.01) closely matching the values for the implicit-bias scores. We then followed Vuletich and Payne's (2019) permutation procedure by (a) shuffling university affiliations, (b) redecomposing IAT scores into within- and between-universities variation, and (c) rerunning models and saving within- and between-groups slopes. Repeating this process 1,000 times, we found that the average ICC from shuffled university affiliations was reduced to near zero (.0001) and found the average between-groups slope ($\bar{\beta}_1$ = 0.27) to be roughly equivalent to the average within-groups slope ($\bar{\beta}_2$ = 0.24). Thus, we were able to produce results mirroring those of Vuletich and Payne simply by adding measurement error to explicit-bias scores. This again suggests that their findings are completely compatible with the idea that implicit bias is a noisily measured individual-level construct.

To be clear, Vuletich and Payne's (2019) permutation result does suggest the existence of some level of systematic group-level variation in implicit bias. But the existence of systematic variation in implicit bias between groups is not in question. Systematic group-level variation exists on virtually any individual-level construct one can think of, so no one would deny that this should also the case for implicit bias. In fact, the existence of at least some level of systematic variation between groups is inherent in the very notion of a group-level test–retest correlation, because without systematic group-level variation, there would be no reason to expect groups with relatively high scores at Time 1 to have relatively high scores at Time 2, and vice versa. Indeed, the simple fact that there is a high test–retest correlation should be enough to convince us of systematic group-level variation; the permutation exercise is superfluous.

A similar confusion pervaded another recent analysis of IAT data by Hehman, Calanchini, Flake, and Leitner (2019). These authors used the same permutation-based approach as Vuletich and Payne (2019) to test whether "the strong explicit–implicit correlations (in racial bias) [parentheses added] at larger level of analysis reflect an artifact of aggregation or, alternately, coherent regional constructs" (p. 1031). Like Vuletich and Payne, the authors observed reduced group-level correlations after shuffling individuals' states of residence and claimed that "these results indicate that true geography matters, and suggest that biases operationalized at the regional level reflects cohesive regional constructs" (p. 1031). Again, however, the existence of high group-level correlations should be enough to convince us that there is at least some level of systematic group-level variation. We do not need permutation to prove this.

High group-level correlations are obviously not purely an "artifact of aggregation": This can be seen clearly if we consider that if there were no group-level variation (i.e., all groups had exactly the same mean), it would be impossible to observe a group-level test–retest or criterion correlation, even if billions of scores were aggregated. Rather, high group-level correlations are an artifact of aggregation plus some level of systematic group-level variation. If we take away systematic group-level variation by randomly assigning group membership, we take away high group-level correlations. But note that—as shown by our bias-of-weekdays example above—the existence of some level of systematic variation at the group level does not mean that there is an important amount of variation at the group level. Very high aggregate-level correlations can in fact represent rather trivial effects, because if sufficient numbers of scores are aggregated within groups, even very small amounts of systematic group-level variation can give rise to extremely high group-level correlations.

An Empirical Test of a Targeted Implicit-Bias-Expulsion Program

So far, we have shown that (a) each of the puzzles regarding implicit bias described by Payne and colleagues (2017a) can be parsimoniously explained as being the expected results of measurement error and aggregation and (b) the counter arguments they have provided against this explanation have been unconvincing. But now we ask: What evidence *could* prove that implicit bias is primarily a feature of environments rather than of individuals?

We propose that one possible test of the model can be based on its tenet that the stability of group means over time is not primarily the result of stability in individual-level scores. This implies that if a particular group appears relatively biased at Time 1, it will also appear relatively biased at Time 2, but this will not be because that group contains specific relatively biased individuals at both time points. Instead, this group-level stability will be due primarily to the fact that "certain contexts encourage discrimination more than others, largely independently of the individual decision makers passing through those contexts." (Vuletich & Payne, 2019, p. 6). Payne and colleagues (2017a) wrote,

> We assume that the average level of implicit bias in a region reflects the average probability of having biased accessible links activated for any given person and any given moment. If the same environmental influences make racially biased links accessible for most people in that context, then the people for whom implicit bias is measured and the people making discriminatory decisions do not need to be the same people. (p. 243)

To test this, we could (in theory!) implement a targeted implicit-bias-based expulsion program in universities. On the basis of implicit-bias scores, we could expel the most biased students from a school. If implicit bias is primarily an individual-level construct tracking stable between-persons differences, this removal of these relatively biased students should lead to a less biased student body on later tests. Moreover, this should also work in reverse; if we expelled the *least* biased students, we would expect a more biased student body on later tests. However, if the bias-of-crowds model is correct, such interventions would have little effect on universities' mean levels of bias, because although these expulsion programs would remove specific individuals from relatively biased and unbiased environments, the environments themselves would stay the same, and according to the model it is the environments, regardless of the specific individuals within them, that are the primary factor in producing stability in implicit bias over time. Expelling students might artifactually reduce aggregate-level mean stability by reducing sample sizes at Time 2, but this effect should be no more extreme than if we randomly chose students to expel from the most biased universities after Time 1.

We used Lai and colleagues' (2016) data on the repeated IAT measurements of students from 18 US universities to test this. First, we restricted the data to students measured on implicit bias at both Time 1 and Time 2. Thus, the total sample size across the 18 universities was 4,841. We then carried out the targeted implicit-bias-based expulsion program. We began by identifying nine high-bias and nine low-bias universities, defined as the universities above or below the university-level median score on Time 1 IAT scores. Iteratively, we then

1. identified the *most* biased student not yet expelled in each high-bias university (the student with the highest Time 1 IAT score) and the *least* biased student not yet expelled in each low-bias university (the student with the lowest Time 1 IAT score);
2. "expelled" those students by excluding their scores from the Time 2 data; and
3. recorded the university-level correlation between Time 1 IAT scores (computed from all students) and Time 2 IAT scores (excluding the expelled students).

We then repeated Steps 1 to 3 50 times. This means that by the 50th iteration, 50 high-bias students (at Time 1) had been expelled at Time 2 from each of the nine high-bias universities, and 50 low-bias students (at Time 1) had been expelled at Time 2 from each of the nine low-bias universities. Importantly, this left the vast majority of most student bodies intact; the average number of students per university was 268.94.[5]

For comparison, we also tested the results of a random expulsion program. This involved a similar process, except that in Step 2 we randomly selected a student from each university to expel. We again iterated each step 50 times, but because of the random variability in results, repeated the entire iteration process 1,000 times.

Figure 6 displays the results, in terms of the resulting university-level test–retest correlations between Time 1 and Time 2 scores. The red line shows that targeted expulsions dramatically reduced the university-level test–retest correlation, and that once 900 total individually targeted students had been expelled, the university-level test–retest correlation became negative ($r = -.21$). By contrast, the blue lines shows that expelling students

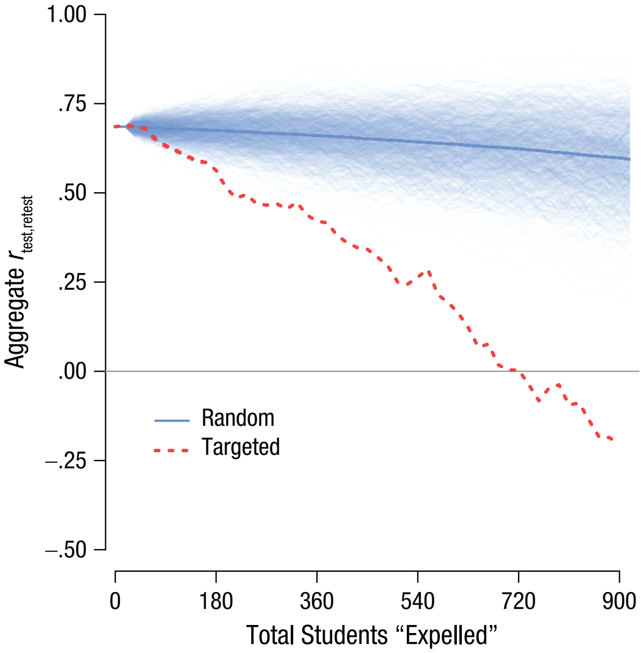

Fig. 6. The effect of targeted and random expulsion. On the basis of Time 1 scores, we expelled the most biased students from the most biased universities and the least biased students from the least biased universities or we expelled students at random. We then tested the effects of these expulsions on aggregate university-level test–retest IAT correlations. The blue shading represents the cloud of unique instances of random expulsion, and the thick blue line is their running average.

randomly had only a slight negative effect on university-level test–retest reliability, although this was variable because of the randomness of the procedure. After 900 randomly selected students had been expelled, the average university-level test–retest correlation remained high ($\bar{r}$ = .60).

Contrary to the claims of Vuletich and Payne (2019), the stability of university-level means in the data gathered by Lai and colleagues (2016) completely relied on the stability of individual-level IAT scores. Universities that were high in bias at Time 1 remained high in bias at Time 2 precisely because they were attended by relatively biased students measured at both time points, and universities that were low in bias at Time 1 remained low in bias at Time 2 precisely because they were attended by relatively unbiased students who were measured at both time points. By "expelling" individually identified students on the basis of their Time 1 IAT scores, the rank-order stability of universities between Time 1 and Time 2 was not only removed but also reversed: The relatively biased universities at Time 1 became the relatively unbiased universities at Time 2, and the relatively unbiased universities at Time 1 became the relatively biased universities at Time 2. This result is extremely difficult to square with the claim that implicit bias is primarily a function of social contexts and not of individuals. If it were truly the structural features of the their environments that primarily give rise to the rank-order stability in universities' mean IAT scores, and implicit bias truly was "a social phenomenon that passes through individuals like "the wave" passes through fans in a stadium" (Vuletich & Payne, 2019, p. 6), it is not at all clear why targeting and expelling a minority of specific individuals from the universities would have so effectively turned those ranks on their head. The only way to explain this result, we believe, is by acknowledging that implicit bias is primarily a noisily measured individual-level construct.

Discussion

We agree with Payne and colleagues (2017a) that social environments can influence individuals' implicit bias. People are not born automatically associating "Black" with "bad" and "White" with "good"; these associations must surely be learned from the environment. And we also do not deny that there appears to be nonnegligible intraindividual variation in implicit bias and that it is plausible, though far from established, that incidental exposure to the kind of structural features of environments discussed by Payne and colleagues influences this variation.

However, simply because a construct is affected by environments does not make it a property of environments. In the present article, we have shown that every piece of evidence put forward in support of Payne and colleagues' bias-of-crowds model is fully compatible with the parsimonious alternate view that implicit bias is primarily an individual-level construct measured with substantial error. Measurement error and aggregation can explain why we see stable group-level means but volatile individual scores and why we see greater correlations with related constructs at the group level than at the individual level. And neither of the counter arguments provided against this alternate view have been convincing. First, the fact of greater internal consistency than test–retest reliability suggests only that factors other than individuals' chronic implicit biases affect implicit-bias test scores, not that those factors are structural features of social environments or that this is somehow unique for implicit bias. Indeed, most individual-level psychological constructs exhibit some level of intraindividual variability and can be affected by incidental features of environments. Second, as we showed above, the fact that group-level slopes are reduced when group membership is randomly shuffled is a natural result of the shuffling process's removing systematic group-level variation; the same thing would occur for virtually any individual-level variable one can imagine.

We therefore see little reason to reconceptualize implicit bias as being a feature of situations rather than individuals, and we believe there are good reasons not to adopt this view. First, as we showed above, each of the core empirical puzzles motivating the model also occur when implicit-bias scores are aggregated within weekdays: Scores are more stable, and correlations with related constructs are greater, at the weekday level than at the individual level. Yet it would be odd to declare implicit bias primarily a feature of weekdays rather than of individuals. Weekdays explain only a tiny fraction of variation in implicit-bias scores, far less than can be explained by individuals' previous scores. To claim that most of the systematic variance in implicit bias is at the weekday level compared with the individual level would be misleading and would represent a misunderstanding of how large aggregate-level correlations can in fact represent relatively trivial amounts of overall variation. We believe the same is true of the state-, country-, or university-level correlations discussed by Payne and colleagues (Payne et al., 2017a; Vuletich & Payne, 2019).

Second, contrary to the implications of the bias-of-crowds model, the rank-order stability of universities' mean levels of implicit bias within Lai and colleagues' (2016) data was completely reliant on the stability of individual-level scores. By removing the most biased students from the most biased universities at Time 1, and the least biased students from the least biased universities at Time 1, the rank order of universities' implicit bias was reversed: The relatively more biased universities at Time 1 became the relatively less biased universities at Time 2, and the relatively less biased universities at Time 1 became the relatively more biased universities at Time 2. And this occurred despite the social context, and the majority of the student bodies, remaining unchanged from Time 1 to Time 2. This result, we believe, cannot be squared with the bias-of-crowds model's claim that mean levels of implicit bias in regions are primarily a function of structural features of their environments, regardless of the specific individuals who happen to inhabit them.

As discussed above, some of the immediate responses to the bias-of-crowds model touched on arguments somewhat similar to ours. However, we believe the present article makes a valuable contribution over and above these initial responses in at least four ways. First, none of the initial responses discussed the crucial point that even very high aggregate-level correlation can represent relatively trivial effects, as our weekday example clearly shows. Second, even the original responses that mentioned the role of measurement error (Kurdi & Banaji, 2017; Rae & Greenwald, 2017) did not discuss the role of ICCs, which, as demonstrated via our simulations, play a key role alongside measurement error and aggregation in producing high aggregate-level correlations from weakly related individual constructs. Third, initial responses were written before Vuletich and Payne's (2019) follow-up article, making ours the first critique of their results and methodology, and thus the first demonstration that their results were completely compatible with implicit bias as a noisily measured individual-level construct. And finally, our targeted implicit-bias-based expulsion program offers a novel and, we believe, convincing demonstration that when put to an empirical test, the bias-of-crowds model fails to find evidentiary support.

What, then, are the implications of recognizing implicit bias as an individual-level construct measured with substantial error for researchers? We believe the field has already made some inroads into grappling with this issue. For example, researchers have begun to acknowledge that studies interested in individual-level differences in implicit bias require a greater focus on measurement than has previously been the norm (Greenwald & Lai, 2020; Kurdi & Banaji, 2017) and that in the absence of substantial advances in the accuracy of measurement tools, single tests of implicit bias cannot and should not be used for the purpose of diagnosing with any certainty the long-term level of bias of any given individual (e.g., Jost, 2019; Kurdi et al., 2019).

Also important, we believe, is for researchers to recognize that although aggregate-level relationships can likely be estimated more accurately than individual-level relationships because of the ability of aggregation to reduce measurement error, these relationships must always be interpreted in the appropriate context. As we showed in our simulations, even when the true relationship between variables exists at the individual level and is relatively weak, it is possible to observe high aggregate-level correlations, if enough individual-level scores are aggregated. So even extremely high aggregate correlations may indicate nothing more than the existence of relatively weak individual-level relationships. Moreover, it is possible for aggregate- and individual-level correlations to be of opposite sign (Simpson, 1951). For example, evidence suggests a positive relationship between implicit weight bias and weight at the country level, but a negative relationship at the individual level (Marini et al., 2013). This can occur for a number of reasons, such as confounding (e.g., culturally varying factors such as the accessibility of fast food for the poor may influence both average weight and attitudes toward people who are overweight) or different causal processes operating at different levels of analysis (e.g., people who are relatively thin or overweight within each country may display a self-serving relative preference for their own body size, whereas increased average weight within a society may cause greater awareness of the health risks of obesity, leading to greater overall stigma). Thus, although researchers may have the ability to measure aggregate-level relationships accurately, they should be careful to always interpret these relationships at the level at which they are observed and to avoid generalizing them to the individual level or overestimating their importance.

As stated at the outset, research on implicit bias has been one of the most meaningful and generative social psychological topics of recent decades and will undoubtedly continue to spur debate as the field continues to refine and develop its understanding and interpretation of its body of evidence. And although we do not ultimately find the bias-of-crowds model to be convincing, we do think it may prove to be a useful signpost for implicit-bias research by forcing researchers to truly grapple with and understand the implications of measurement error and aggregation and, perhaps, by renewing interest in studying the factors that underlie intraindividual variation in implicit bias.

Transparency

Action Editor: Laura A. King
Editor: Laura A. King
Declaration of Conflicting Interests
The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

ORCID iDs

Paul Connor https://orcid.org/0000-0002-4995-4679
Ellen R. K. Evers https://orcid.org/0000-0002-8667-3083

Acknowledgments

We thank Leif Nelson, Michael Zyphur, Stephen Antonoplis, Everett Wetchler, and Smriti Mehta for their constructive feedback on early drafts of the manuscript, and Alex Kogan for sharing his insight into the nature of individual-level measurement error and aggregation.

Notes

[1] All data and code used for all analyses and simulations is available on our Open Science Framework page: (https://osf.io/tj8u6/).

[2] In fact there are 1,939 identifiable U.S. cases with matching User IDs measured more than once on implicit bias, but a number of these do not appear to be the same person across measurement occasions. For example, multiple cases with the same User ID had self-reported age (taking into account measurement date), gender, or race that did not match up across measurement occasions, so we chose to exclude these cases from calculations.

[3] For the criterion correlation the choice to use test rather than retest is arbitrary and does not affect overall results.

[4] The D score, which is analogous to Cohen's *d* at the participant level, is calculated by dividing the within-person difference by the standard deviation of the practice and critical blocks combined.

[5] Two universities had less than 100 students sampled at both time points, Iona College (n = 93) and University of Virginia at Wise (n = 82).

References

Amodio, D. M., & Mendoza, S. A. (2010). Implicit intergroup bias: Cognitive, affective, and motivational underpinnings. In B. Gawronski & B. K. Payne (Eds.), *Handbook of implicit social cognition: Measurement, theory, and applications* (pp. 353–374). New York, NY: The Guilford Press.

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

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis in the behavioral sciences* (3rd ed.). Mahwah, NJ: Erlbaum.

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

Devine, P. G., Plant, E. A., Amodio, D. M., Harmon-Jones, E., & Vance, S. L. (2002). The regulation of explicit and implicit race bias: The role of motivations to respond without prejudice. *Journal of Personality and Social Psychology*, *82*, 835–848.

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*(6), 1013–1027.

Furr, R. M., & Bacharach, V. R. (2013). *Psychometrics: An introduction*. Los Angeles, CA: Sage.

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.

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. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology*, *74*, 1464–1480.

Greenwald, A. G., Nosek, B. A., & Banaji, M. R. (2003). Understanding and using the implicit association test: I. An improved scoring algorithm. *Journal of Personality and Social Psychology*, *85*, 197–216.

Greenwald, A. G., Poehlman, T. A., Uhlmann, E. L., & Banaji, M. R. (2009). Understanding and using the Implicit Association Test: III. Meta-analysis of predictive validity. *Journal of Personality and Social Psychology*, *97*, 17–41.

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.

Johnson, T. J., Hickey, R. W., Switzer, G. E., Miller, E., Winger, D. G., Nguyen, M., . . . Hausmann, L. R. (2016). The impact of cognitive stressors in the emergency department on physician implicit racial bias. *Academic Emergency Medicine*, *23*, 297–305.

Jost, J. T. (2019). The IAT is dead, long live the IAT: Context-sensitive measures of implicit attitudes are indispensable to social and political psychology. *Current Directions in Psychological Science*, *28*(1), 10–19.

Kurdi, B., & Banaji, M. R. (2017). Reports of the death of the individual difference approach to implicit social cognition may be greatly exaggerated: A commentary on Payne, Vuletich, and Lundberg. *Psychological Inquiry*, *28*, 281–287.

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

Lai, C. K., Hoffman, K. M., & Nosek, B. A. (2013). Reducing implicit prejudice. *Social and Personality Psychology Compass*, *7*, 315–330.

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

Marini, M., Sriram, N., Schnabel, K., Maliszewski, N., Devos, T., Ekehammar, B., . . . Schnall, S. (2013). Overweight people have low levels of implicit weight bias, but overweight nations have high levels of implicit weight bias. *PLOS ONE*, *8*(12), Article e83543. doi:10.1371/journal.pone.0083543

Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining R^2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, *4*(2), 133–142.

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., Vuletich, H. A., & Brown-Iannuzzi, J. L. (2019). Historical roots of implicit bias in slavery. *Proceedings of the National Academy of Sciences, USA*, *116*(24), 11693–11698.

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

Payne, B. K., Vuletich, H. A., & Lundberg, K. B. (2017b). Flipping the script on implicit bias research with the bias of crowds. *Psychological Inquiry*, *28*, 306–311.

Rae, J. R., & Greenwald, A. G. (2017). Persons or situations? Individual differences explain variance in aggregated implicit race attitudes. *Psychological Inquiry*, *28*, 297–300.

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

R Core Team. (2019). R: A language and environment for statistical computing (Version 3.6.1) [Computer software]. Vienna, Austria: R Foundation for Statistical Computing. Available from https://www.R-project.org/

Schwarz, N., Strack, F., Kommer, D., & Wagner, D. (1987). Soccer, rooms, and the quality of your life: Mood effects on judgments of satisfaction with life in general and with specific domains. *European Journal of Social Psychology*, *17*, 69–79.

Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society B: Methodological*, *13*, 238–241.

Singal, J. (2017 January). Psychology's favorite tool for measuring racism isn't up to the job. *New York Magazine*. Retrieved from http://nymag.com/scienceofus/2017/01/psychologys-racism-measuring-tool-isnt-up-to-the-job.html

Vuletich, H. A., & Payne, B. K. (2019). Stability and change in implicit bias. *Psychological Science*, *30*, 854–862.

Weir, K. (2016). Policing in black and white. *Monitor on Psychology*, *47*(11), 36–43.

Xu, K., Nosek, B., & Greenwald, A. (2014). Data from the race implicit association test on the Project Implicit demo website. *Journal of Open Psychology Data*, *2*(1), Article p3. doi:10.5334/jopd.ac