Stability and Change in Implicit Bias

Heidi A. Vuletich and B. Keith Payne
Department of Psychology and Neuroscience, University of North Carolina at Chapel Hill

Abstract
Can implicit bias be changed? In a recent longitudinal study, Lai and colleagues (2016, Study 2) compared nine interventions intended to reduce racial bias across 18 university campuses. Although all interventions changed participants’ bias on an immediate test, none were effective after a delay. This study has been interpreted as strong evidence that implicit biases are difficult to change. We revisited Lai et al.’s study to test whether the stability observed reflected persistent individual attitudes or stable environments. Our reanalysis (N = 4,842) indicates that individual biases did not return to preexisting levels. Instead, campus means returned to preexisting campus means, whereas individual scores fluctuated mostly randomly. Campus means were predicted by markers of structural inequality. Our results are consistent with the theory that implicit bias reflects biases in the environment rather than individual dispositions. This conclusion is nearly the opposite of the original interpretation: Although social environments are stable, individual implicit biases are ephemeral.

Keywords
implicit bias, bias of crowds, implicit attitudes, IAT, prejudice, open data

Received 5/17/18; Revision accepted 1/22/19

Are implicit biases indelible attitudes ingrained through years of exposure to society’s prejudice? Or are they momentary patterns of semantic activation, as changeable as a stream of thought? Questions of stability and change are important for understanding the nature of implicit bias and for prospects of reducing unintended discrimination.

Implicit biases are automatic associations with social groups (Fazio & Olson, 2003; Gawronski & Bodenhausen, 2006). They are considered biases because different associations are linked via social stereotypes to different groups. Even though an association between a group and stereotyped concepts does not imply intentional animosity, implicit biases have been theorized to be a source of (perhaps unintentional) discriminatory treatment (Bargh, 1999; Devine, 1989; Greenwald & Banaji, 1995). Early theories assumed that implicit biases were difficult to change (Bargh, 1999; Devine, 1989; Wilson, Lindsey, & Schooler, 2000). Later research, however, found that implicit tests were malleable in response to mental imagery, shifts in attention and goals, salient exemplars, and other interventions (Blair, 2002; Dasgupta, 2013; Lai, Hoffman, & Nosek, 2013; Payne & Gawronski, 2010). A meta-analysis of procedures intended to change implicit bias found evidence that scores were malleable when measured on immediate tests, but few studies assessed changes after a delay (Forscher et al., 2017).

The most comprehensive evidence of stability comes from a large-scale experimental study (Lai et al., 2016). Nine interventions were examined using a longitudinal design. Data were collected from 18 university campuses. This study included a pretest measure of implicit race bias, followed by an intervention phase with an immediate posttest, and a follow-up measure completed up to a few days later. The study had a large sample, experimental control, and a longitudinal design, making it the highest quality test of stability and change currently available.

Results of the study indicated that all nine interventions were effective on the immediate test, yet none of the interventions produced a lasting effect after 1 to 2
days. These effects have been interpreted as strong evidence that implicit biases are rigid. As Lai and colleagues concluded, “These findings are a testament to how the mind’s prejudices remain steadfast in the face of efforts to change them” (p. 1014). According to this view, individuals’ attitudes snap stubbornly back to their baseline after a short delay.

In this article, we suggest an alternative interpretation of the findings of Lai and colleagues (2016). Rather than reflecting rigid attitudes, the absence of intervention effects following a delay may instead reflect highly malleable individual attitudes, constrained by the stability of social environments. Our reasoning is based on the bias-of-crowds model of implicit bias (Payne, Vuletich, & Lundberg, 2017).

The bias-of-crowds model posits that implicit bias is driven by the cognitive accessibility of concepts linked to social categories. Accessibility refers to the likelihood that a piece of information will be retrieved and used in later processing (Fazio & Williams, 1986; Higgins, 1996; Srull & Wyer, 1979). The accessibility of associations can vary both chronically (as a feature of the person) and situationally (as an aspect of the environment). However, reviews of the literature suggest that measures of implicit bias tend to be temporally unstable (Gawronski, Morrison, Phillips, & Galdi, 2017) and produce small correlations with outcomes as individual-difference measures (Cameron, Brown-Iannuzzi, & Payne, 2012; Greenwald, Poehlman, Uhlmann, & Banaji, 2009; Oswald, Mitchell, Blanton, Jaccard, & Tetlock, 2013). In contrast, aggregate levels of implicit bias across cities, counties, or nations tend to be highly stable and produce large associations with aggregate-level racial disparities (Hehman, Flake, & Calanchini, 2018; Leitner, Hehman, Ayduk, & Mendoza-Denton, 2016; Orchard & Price, 2017). This evidence suggests that accessibility may vary systematically as a function of situations rather than as a result of individual dispositions.

The bias-of-crowds model also posits that implicit biases are largely transient at the level of individuals. For an individual, the concepts most accessible at any moment depend on countless factors, from shared cultural stereotypes to unshared experiences such as recent media exposures and fleeting thoughts. However, when independent observations are aggregated, they function like the wisdom-of-crowds phenomenon in which partial knowledge distributed among many individuals gives rise to stable and accurate aggregate estimates (Surowiecki, 2004). When individual measures of implicit bias are aggregated, the randomly distributed transient influences will be averaged away, whereas shared environmental influences will be sharpened. As a result, the average level of implicit bias in an environment converges on an accurate estimate of the level of cultural stereotypes and structural inequalities in that environment.

Evidence consistent with the structural-inequality account was reported in a study linking geographical differences in average levels of implicit bias to geographical differences in historical slavery (Payne, Vuletich, & Brown-Iannuzzi, in press). The study compared the proportion of the population enslaved, according to the 1860 census, with county-level implicit race bias in more than 1,400 counties from the Project Implicit database (Xu, Nosek, & Greenwald, 2014). Much research suggests that economic dependence on slavery motivated a range of cultural, legal, economic, and ideological reactions aimed at justifying slavery and, later, maintaining the racial hierarchy (Acharya, Blackwell, & Sen, 2018; Anderson, 2016; Rothstein, 2017). Those structural inequalities were hypothesized to be associated with present-day implicit bias. As we predicted, states and counties that were more dependent on enslaved labor in 1860 displayed higher levels of implicit bias more than 150 years later. The association between slave populations and implicit race bias was partially mediated by markers of structural inequalities, including present-day residential segregation and inter-generational mobility.

These findings suggest that aggregate levels of implicit bias reflect structural inequalities in the environment, consistent with the bias-of-crowds model. Applying the model to understand the results of Lai and colleagues suggests that the lack of intervention effects after a delay may not reflect stable individual attitudes, as originally concluded. In a reanalysis of the data, we hypothesized that individual implicit-bias scores would not return systematically to their preintervention level. Instead, we expected individual scores to fluctuate from one time point to another, producing low stability at the individual level. In contrast, we expected the campus-level averages for the posttest to return to levels similar to those at the campus-level pretest. To test whether these mean levels reflect structural inequalities, we examined the association between campus-level means and several markers of structural inequalities.

If our interpretation is correct, it would substantially change the conclusion from Lai et al.’s study. Far from being rigid attitudes, implicit biases could be like the weather: If you want it to change, simply wait a day or two. Average environmental levels of bias, in contrast, may be more like climate. Changes are slow, but even small changes may have immense consequences.

**Method**

To test our interpretation, we reanalyzed the data from Study 2 by Lai et al. (2016; data can be found at
The third marker of structural inequality was a measure of campus-specific social mobility. The data were retrieved from a large-scale study that estimated social mobility for nearly all universities in the United States (Chetty, Friedman, Saez, Turner, & Yagan, 2017). This measure reflects the percentage of students whose parents occupied the poorest income quintile and who made it to the top quintile in adulthood. Data are based on the 2014 earnings of former college students who attended college during the early 2000s. Although this measure is not race specific, the association between race and socioeconomic status means that low social mobility has a disproportionate impact on members of minority groups.

We collected data on students’ SAT scores at each university (we used the 75th percentile because the median was not available) and each university’s admissions selectivity (number admitted divided by the number of applications) as control variables. The data, collected in 2017, were accessed from the U.S. Department of Education’s Integrated Postsecondary Education Data System (National Center for Education Statistics, 2019). These variables were chosen to help separate measures of institutional inequality from overall selectivity.

**Results**

As reported by Lai et al., average IAT scores decreased from Time 1 to Time 2 and then rebounded at Time 3. Table 1 displays the descriptive statistics for each campus. To better understand sources of stability and change, we tested three primary hypotheses. The first hypothesis was that individual-level implicit-bias scores would show low stability over time. The second hypothesis was that campus-level means, in contrast, would show high stability. The third hypothesis was that campus-level implicit bias would be associated with markers of structural inequalities.

As displayed in Figure 1, the person-level test–retest correlation between Time 1 and Time 3 was $r(4839) = .25$, $p < .001$, bootstrapped 95% confidence interval (CI) = [.22, .27]. A possible reason for the low test–retest stability is that the interventions may have affected individuals in heterogeneous ways, thus changing their rank orders over time. To test this possibility, we examined the stability of scores among participants in the control condition, who received no intervention. The association between Time 1 and Time 3 scores was identical to those in the larger sample, $r(462) = .25$, $p < .001$, bootstrapped 95% CI = [.15, .34]. The low stability observed here contrasts with the interpretation that individual biases are difficult to change, because individual scores did not generally return to pretest levels. Individual scores changed a great deal, although not systematically.
Next, we examined the role of environments in creating stable means. To the extent that stability is found in local environments, we expected Time 3 campus means to be predicted by Time 1 campus means. As displayed in Figure 2, the bivariate association was strong, $r(16) = .72$, $p < .001$, bootstrapped 95% CI = [.47, .99].

We further investigated the role of individuals and environments simultaneously by estimating a multilevel model to account for the nested structure of the data. Individual scores (Level 1) were nested within university (Level 2). IAT scores at Time 1 were campus-mean centered to separate within- and between-campus effects and were used to predict raw Time 3 scores (Raudenbush & Bryk, 2002). Mean-centering scores by campus yielded an unbiased estimate of the individual-level association between Time 1 scores and Time 3 scores, holding campus effects constant. To obtain an estimate of the campus effect, we included site means at Level 2 as a predictor of Time 3 intercepts. The equations for the Level 1 (person-level) model were as follows:

$$IAT \text{ Time 3} = \beta_{0j} + \beta_{1j} \text{ IAT}_{mc} + r_{ij}$$

$$r_{ij} \sim N(0, \sigma^2).$$

The equations for the Level 2 (campus level) model were as follows:

$$\beta_{0j} = \gamma_{00} + \gamma_{01} \text{ site}_{i} \text{ IAT}_{j} + \mu_{0j}$$

$$\mu_{0j} \sim N(0, \tau_{00})$$

$$\beta_{1j} = \gamma_{11},$$

where $\gamma$ is the fixed effect, $r$ is the residual, $\mu$ is the random effect, $N$ is the total number of units at Level 1,
$\sigma^2$ is the Level 1 residual variance, and $\tau_{0j}$ is the variance of the random effect. T1 refers to Time 1, T3 to Time 3, cmc to campus-mean centered, $i$ to individual, and $j$ to campus.

The university-level means at Time 1 were strongly predictive of means at Time 3, $\gamma_{01} = 0.87$, $p < .001$, 95% CI = [0.44, 1.29]. This effect was much larger than the stability of individual scores across time, $\gamma_{10} = 0.25$, $p < .001$, 95% CI = [0.22, 0.27]. As can be seen by the confidence intervals, the coefficient for campus means was significantly different from the coefficient for individual scores. Given these comparisons, the high stability in means observed by Lai et al. does not appear to result primarily from individuals reverting to their original attitudes following the intervention. Instead, the stability appears to result from university means returning with high fidelity to earlier university means.

A critical difference between subject-level and campus-level scores is that campus-level scores are aggregated across many individuals. The intraclass correlation coefficient was .01, meaning that campuses accounted for approximately 1% of the variability in individual scores. This small association is consistent with our reasoning that the effect of contexts on individual implicit-test scores may be small and fleeting, and the context-based signal is revealed only when noise is reduced by aggregation across individuals.

Did this aggregation reveal the level of implicit bias on each campus or spuriously create the observed relationship? To examine the effect of aggregation alone, we randomly reassigned all person-level observations to 18 new nominal groups of the same size as those in the original sample, using sampling without replacement (for a similar approach, see Hehman, Calanchini, Flake, & Leitner, 2018). Then, we estimated the multilevel model and repeated this randomization and analysis procedure a total of 100 times. If the stable means that we observed in the original sample were caused by mere aggregation, we would expect the means of the randomized groups to also be stable. But if aggregation instead reveals systematic variance at the group level, then randomizing the groups should reduce the group-level effect. In fact, the campus-level coefficient was greatly reduced, average $\gamma_{01} = 0.26$, 95% CI = [0.21, 0.31]. Interestingly, this coefficient approximated the same value as the person-level effect (0.25). 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.

Our final analysis was intended to shed light on the nature of the campus-specific signal. We hypothesized that implicit bias would be associated with markers of structural inequalities. As can be seen in Table 2, Time 1 implicit bias was significantly correlated with the three structural-inequality measures. IAT means were higher on campuses with a Confederate monument displayed, $\tau(16) = .64$, $p = .005$, bootstrapped 95% CI = [.39, .89], and lower on campuses with more faculty diversity, $\tau(16) = -.48$, $p = .043$, bootstrapped 95% CI = [-.91, -.12], and with greater social mobility, $\tau(16) = -.61$, $p = .007$, bootstrapped 95% CI = [-1.00, -.25].

To create a more robust measure of structural inequalities, we constructed a composite index by reverse-scoring faculty diversity and mobility so that higher values reflect less diversity and poorer mobility. Then, we averaged the (standardized) variables into a structural-inequality index (Cronbach’s $\alpha = .71$). The association between Time 1 implicit bias and the structural-inequality index was strong, $\tau(16) = .73$, $p = .001$, bootstrapped 95% CI = [.47, 1.00] (see Fig. 3). When we controlled for campus SAT scores and admissions selectivity, the structural-inequality coefficient remained significant and large, $\beta = 0.67$, $p = .006$, 95% CI = [0.65, 0.69].

**Discussion**

Evidence that the effects of implicit-bias interventions are short-lived has been interpreted as showing that individuals’ implicit biases are difficult to change (Lai et al., 2016). We reanalyzed the data by considering both individuals and environments as potential sources of stability. Our analyses suggest that at the individual level, implicit biases were far from permanent. Average bias at the follow-up test, however, was well predicted by the preexisting average at each university. The preexisting level of implicit bias, in turn, was associated with measures of structural inequality. Consistent with the bias-of-crowds model, fleeting biases at the individual level revealed stable and meaningful estimates of environmental bias when aggregated.
From a traitlike view of implicit bias, the combination of low stability in individual attitudes with high stability in means is puzzling. But the bias-of-crowds model provides a natural explanation for the return to preexisting averages after a delay, despite low stability in individual attitudes.

The strong campus-level stability is unlikely to be an artifact of aggregation. The same degree of aggregation did not produce high stability when individuals were randomly reshuffled into nominal groups. Instead, aggregation revealed meaningful differences among campus contexts. Those differences were systematically associated with measures of structural inequality, providing further evidence that campus-level biases are meaningful.

Aggregating repeated observations of the same subjects could potentially produce stable estimates of individual attitudes by revealing chronic accessibility effects. In practice, stable individual scores may require up to a dozen observations per subject (Kurdi & Banaji, 2017). Our core claim in this article is that individual-level implicit bias is not a rigid attitude. The need for many observations to obtain a stable individual measure is consistent with that claim, because it suggests that implicit bias must not be as rigid as was once thought.

Our reinterpretation has fundamental implications for the nature of implicit bias. Most theories of implicit bias assume that it is a property of individuals, such as an attitude, belief, or trait. The context-based perspective, in contrast, suggests that implicit bias is a social phenomenon that passes through individuals like “the wave” passes through fans in a stadium. Rather than a property of individuals, it may more properly be considered a property of social contexts. Most theories assume that a person’s implicit bias is difficult to change. The context-based view, in contrast, suggests that a person’s level of implicit bias is transient and can change as often as the context changes.

Most theories assume that individuals with high levels of bias are substantially more likely to discriminate than those with low levels of bias. The context-based view, in contrast, suggests that certain contexts encourage discrimination more than others, largely independently of the individual decision makers passing through those contexts. In these ways, the context-based view advanced by the bias-of-crowds model reverses core theoretical assumptions about the nature of implicit bias.

Our reinterpretation also has practical implications for reducing discrimination. Namely, changing the social environment may be more effective than changing individual attitudes. Environmental interventions might take either of two forms. Temporary interventions, such as the ones tested by Lai and colleagues,
may be powerful if they are targeted at the time when important decisions are being made. For example, cuing decision makers to think about counterstereotypical thoughts or affirming egalitarian values immediately before making hiring or admissions decisions may reduce unintended bias in those decisions. Rather than changing attitudes, this strategy may modify the concepts that are most accessible in the decision situation. A second approach is to change social environments in more lasting ways. Our findings suggest the hypothesis that removing environmental cues of inequality, such as Confederate monuments, may reduce aggregate implicit bias. Likewise, increasing faculty diversity at universities, or diversity in an organization’s leadership more generally, may produce sustained changes in institutional bias.

These ideas remain to be tested, because this correlational design could not establish the causal role of the campus contexts. Nor can it rule out selection effects. However, if the effects were primarily driven by attributes of the students that attracted them to particular colleges, then we would expect to see relatively large person-level effects. Nonetheless, research that randomly assigns individuals to different contexts is an important next step for understanding the causal role of contexts in cuing bias. Another limitation of this study is that, despite the large sample, having only 18 campuses provides relatively low power for comparisons at the campus level.

In the reported research, we measured implicit bias using the IAT, which is only one of multiple implicit tests available and has psychometric limitations (Fiedler, Messner, & Bluemke, 2006). Future research should examine the relative importance of person and context effects using other measures, such as the affect-misattribution procedure (Payne, Cheng, Govorun, & Stewart, 2005) or sequential priming (Fazio, Jackson, Dunton, & Williams, 1995).

Despite these limitations, our research suggests that a core claim in much research on implicit bias may need to be revised. Far from being a rigid attitude, implicit bias is highly transient at the individual level but stable for social contexts. If these findings are confirmed in future research, they suggest that the source of stable implicit bias—and the opportunity for change—is to be found in the places and people around us.

**Action Editor**

Brent W. Roberts served as action editor for this article.

**Author Contributions**

Both authors analyzed the data, wrote the manuscript, and approved the final manuscript for submission.

**ORCID iD**

Heidi A. Vuletich [ORCID iD](https://orcid.org/0000-0002-8614-4333)

**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.

**Funding**

This work was funded in part by a National Science Foundation Graduate Fellowship and the Paul and Daisy Soros Fellowships for New Americans (both awarded to H. A. Vuletich) and a National Science Foundation Research Grant (awarded to B. K. Payne).

**Open Practices**

All data have been made publicly available via the Open Science Framework and can be accessed at osf.io/ng9k8. The complete Open Practices Disclosure for this article can be found at http://journals.sagepub.com/doi/suppl/10.1177/0956797619844270. This article has received the badge for Open Data. More information about the Open Practices badges can be found at http://www.psychologicalscience.org/publications/badges.

**Notes**

1. The original article described 17 sites because the researchers combined two campuses of the University of Virginia, but we separated these because they were in different counties.

2. Social-mobility statistics were not available for Tulane University. However, county-level mobility data were available, and the two were highly correlated across the other campuses, $r(15) = .70, p = .002$. We therefore used county-level mobility to impute campus-level mobility for Tulane.

3. Data and associated code are available at osf.io/ng9k8.

**References**


from https://www.chronicle.com/article/We-re-Tracking-Confederate/240967


