Journal of Experimental Social Psychology 72 (2017) 133–146

Contents lists available at ScienceDirect

Journal of Experimental Social Psychology

journal homepage: www.elsevier.com/locate/jesp

Breaking the prejudice habit: Mechanisms, timecourse, and longevity☆,☆☆,★

Patrick S. Forscher*, Chelsea Mitamura, Emily L. Dix, William T.L. Cox, Patricia G. Devine*

Department of Psychology, University of Wisconsin, Madison, United States

ARTICLE INFO

Keywords:
Intervention
Implicit bias
Social cognition
Replication

ABSTRACT

The prejudice habit-breaking intervention (Devine, Forscher, Austin, & Cox, 2012) and its offshoots (e.g., Carnes et al., 2015) have shown promise in effecting long-term change in key outcomes related to intergroup bias, including increases in awareness, concern about discrimination, and, in one study, long-term decreases in implicit bias. This intervention is based on the premise that unintentional bias is like a habit that can be broken with sufficient motivation, awareness, and effort. We conducted replication of the original habit-breaking intervention experiment in a sample more than three times the size of the original (N = 292). We also measured all outcomes every other day for 14 days and measured potential mechanisms for the intervention's effects. Consistent with previous results, the habit-breaking intervention produced a change in concern that endured two weeks post-intervention. These effects were associated with increased sensitivity to the biases of others and an increased tendency to label biases as wrong. Contrasting with the original work, both control and intervention participants decreased in implicit bias, and the effects of the habit-breaking intervention on awareness declined in the second week of the study. In a subsample recruited two years later, intervention participants were more likely than control participants to object on a public online forum to an essay endorsing racial stereotyping. Our results suggest that the habit-breaking intervention produces enduring changes in peoples' knowledge of and beliefs about race-related issues, and we argue that these changes are even more important for promoting long-term behavioral change than are changes in implicit bias.

Intergroup inequality is a ubiquitous problem, with minorities facing disparities in a variety of consequential domains, from the allocation of medical care (Williams, Neighbors, & Jackson, 2003) to hiring (Bertrand & Mullainathan, 2004). One potential contributor to these problems is implicit bias. Theorized to influence behavior despite countervailing intentions (Devine, 1989), implicit bias has attracted attention precisely because it has the potential to cause otherwise fair-minded people to be unwittingly complicit in the perpetuation of inequality. The specter of unintentional discrimination has inspired widespread calls from researchers, scholars, and public policy officials to develop effective interventions to reduce and eliminate the negative effects of unintentional bias (e.g., Fiske, 1998; Smedley, Stith, & Nelson, 2003). Revealing the scope of the response to this call, a recent meta-analysis uncovered 573 experiments testing methods to change implicit bias (Forscher et al., 2017). Though many interventions reduced implicit bias, very few (6.6% of the meta-analytic samples) have been tested over time. Solving social problems requires interventions that produce changes that endure.

The prejudice habit-breaking intervention contrasts with many of these other interventions in that it was explicitly developed to produce enduring change (Devine et al., 2012; see also Carnes et al., 2015). The habit-breaking intervention is based on the prejudice habit model (Devine, 1989), which proposes that enduring change in biases, such as implicit bias, that occur unintentionally can be achieved by treating unintentional bias as an unwanted habit that can be broken through a combination of motivation, awareness, and effort. This multifaceted intervention was designed to address a number of common stumbling blocks on the path to change. Specifically, although many people feel motivated to overcome biases in their behavior (Devine, Monteith, Zuwerink, & Elliot, 1991; Monteith, 1993; Monteith, Ashburn-Nardo, Voils, & Czopp, 2002; Plant & Devine, 1998), they are not always aware of their biases (Monteith, Voils, & Ashburn-Nardo, 2001), nor do they

☆ Conceived research: Forscher, Devine; Designed research: Forscher, Mitamura, Devine; Coordinated data collection: Forscher, Mitamura; Coded free response data: Forscher, Mitamura, Dix; Analyzed data: Forscher, Dix; Wrote paper: all authors; Revised paper: all authors.

☆☆ We'd like to thank Maha Baalbaki, Brittany Drifka, Collin Eckberg, Hannah Kimyon, Keith Knutson, Alyssa Law, Tianrui Li, Mary Martinco, Ryan Massopust, Kelly Nance, Nicole Sather, and Zach Wittrock for their help collecting data for this study.

★ Preparation of this article was supported by NIH grant 5R01GM111002-02, a Wisconsin Alumni Research Foundation Professorship awarded to the last author, and a National Science Foundation Graduate Research Fellowship (DGE-1256259) awarded to the third author. Data and materials for this project can be found at https://osf.io/a3c8h/.

* Corresponding authors.
E-mail addresses: schnarrd@gmail.com (P.S. Forscher), pgdevine@wisc.edu (P.G. Devine).

http://dx.doi.org/10.1016/j.jesp.2017.04.009
Received 25 September 2016; Accepted 26 April 2017
Available online 11 May 2017
0022-1031/ Published by Elsevier Inc.

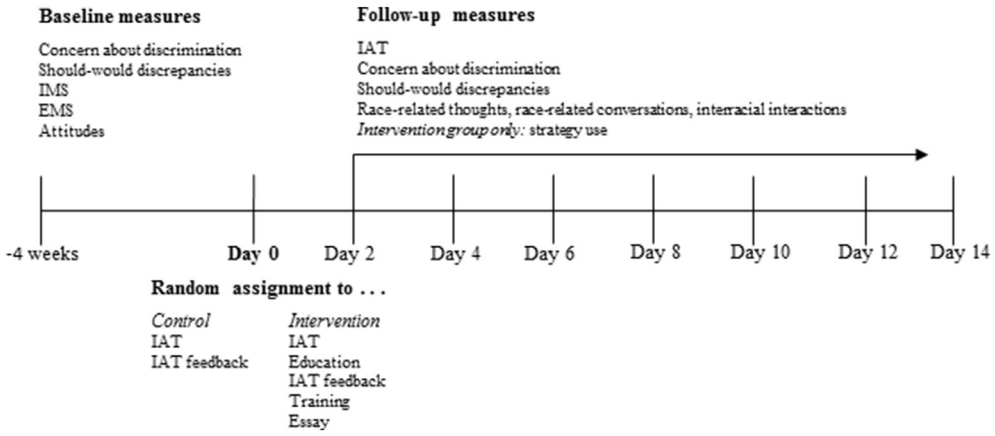

Fig. 1. Study timeline.

always know how to productively channel their motivation into behavior that will help overcome bias (Apfelbaum, Sommers, & Norton, 2008; Norton, Sommers, Apfelbaum, Pura, & Ariely, 2006). In some cases, the threat of behaving with bias may even lead people to avoid members of minority groups in an effort to prevent the possibility of biased behavior and the negative feelings that follow (Plant & Devine, 2003; Stephan & Stephan, 1985).

Using a semi-interactive slide show, the prejudice habit-breaking intervention navigates people around these stumbling blocks by providing education about the existence, origins, and consequences of unintentional bias, and teaching them evidence-based strategies to overcome bias. This presentation provokes awareness by giving participants feedback about their own level of implicit bias, as measured by the Implicit Association Test (IAT; Greenwald, McGhee, & Schwartz, 1998). The presentation then teaches them how implicit bias can lead to unintentional but consequential discriminatory behavior, leading to negative consequences for racial minorities. Finally, it provides recipients with evidence-based, cognitive strategies that, if practiced, can lead to bias reduction (i.e., stereotype replacement, perspective taking, individuation, counterstereotypic imaging, and increasing opportunities for contact). These strategies give participants productive ways of channeling their behavior into effective solutions that they can implement independently to reduce their bias over time. In a randomized controlled trial, Devine and colleagues demonstrated that the habit-breaking intervention produced long-term changes in key outcomes, including reduced IAT bias, increased concern about discrimination, and greater reported beliefs that there could be bias present in their thoughts, feelings, and behaviors. These changes endured two months following the intervention.

The long-term effects of the habit-breaking intervention were both exciting and encouraging, suggesting that this intervention might be an effective tool in our efforts to more broadly address sources of persistent social inequality. Before it can fulfill this role, however, we must know how the effects occur, whether they replicate, whether they last longer than two months, and whether they generalize to consequential behaviors. The present work was developed to address these questions. Specifically, we conducted a new experiment to replicate the habit-breaking intervention's effects on implicit bias, concern, and awareness in a larger sample of participants. We further assessed some potential mechanisms of the intervention's effects and examined the extent to which the intervention affected behavior two years later.

With these objectives in mind, we evaluated the effects of the habit-breaking intervention in a two-phase design. In the first phase, we randomly assigned participants to intervention and control conditions and measured a set of outcomes every other day for two weeks. Specifically, every two days, we collected measures of: (1) implicit bias, (2) concern about discrimination, (3) discrepancies between participants' standards for and beliefs about their interracial behavior, and (4) several potential mechanisms for the intervention's effects: the quantity and content of race-related thoughts, race-related conversations, and cross-race interactions with Black people. Participants in the intervention condition were also asked whether and how they used the bias-reduction strategies learned during the intervention. We chose this frequent assessment schedule so that we could obtain stable estimates of the intervention's effects, assess trajectories of change, and assess the relationships between change processes. In the second phase, we invited these same participants to complete an ostensibly unrelated study two years later. Phase 2 contained key measures from Phase 1, as well as three behaviors that could plausibly be affected by the habit-breaking intervention.

1. Phase 1

1.1. Method

All materials, data, and supplemental analyses are publicly available at https://osf.io/a3c8h/.

1.1.1. Participants

In Phase 1, 302 non-Black students in Introductory Psychology were randomly assigned to intervention (N = 138) and control (N = 164) conditions. We aimed to obtain a much larger sample size than Devine and colleagues by recruiting as many participants as possible over the course of two semesters. Of the initial 302 participants, four from the control condition were eliminated prior to analysis because they mistakenly received follow-ups designed for the intervention group (referencing the bias-reduction strategies). An additional six participants were eliminated because, on multiple occasions our measure of implicit bias, they had mean reaction times below 400 ms and/or mean accuracies below 70%, suggesting a lack of task attention. These eliminations resulted in a total of 292 participants (136 intervention, 156 control; 68% female, 67% White, 25% Asian) who were eligible for analysis, more than three times the sample of 91 used by Devine et al. (2012).

1.1.2. Procedure

As shown in Fig. 1, our procedure was highly similar to that used by Devine et al. (2012), with six differences that are shown in Table 1 and described in more detail below. At the beginning of the semester, we measured baseline concern and discrepancies between standards and beliefs as part of a large online survey. The online survey also contained measures of attitudes toward Black people and the internal and external motivations for responding without prejudice (IMS and EMS; Plant & Devine, 1998), which we included to ensure equivalence between experimental conditions on these dimensions at baseline. We allowed participants to enroll in the study between two weeks and two months after the completion of the large online survey.

Participants completed the first session in the lab in groups of one to

Table 1
Differences in procedure between Devine et al. (2012) and the current study.

Procedural element	Devine et al. (2012)	Current study	Justification
Sequencing of IAT feedback	After IAT and before slideshow	After education section of slideshow	Feedback makes more sense once a person knows about what IAT measures
Essay intervention's benefits	Not administered	Administered to intervention participants after slideshow	Enhance intervention's effects through self-persuasion; use as measure
Study duration	Two months	Fourteen days	Focus on period soon after intervention
Number of follow-ups	Two	Seven	Increase precision; understand timecourse
Frequency of follow-ups	Every month	Every other day	Understand changes that occur rapidly
Mechanism measures	None	Race-related thoughts & conversations; interracial interactions	Understand mechanisms of intervention

six. Each group was randomly assigned to condition.[1] All participants were then asked to complete our measure of implicit bias, the Black-White evaluative Implicit Association Test (IAT). After the IAT, control participants received feedback about their IAT scores, completed a measure of affect that is not discussed further, and were dismissed. The IAT feedback and affect measure were included to match the design of Devine et al. (2012), who sought to establish that feedback alone is insufficient to achieve the habit-breaking intervention's effects. Participants in the intervention condition completed the narrated, semi-interactive slideshow that constitutes the prejudice habit-breaking intervention. Embedded in this slideshow was feedback about their IAT score and the same affect measure given to control participants. To encourage participants to pay close attention to the content of intervention, they were told that we were considering adapting the intervention for use with high school students. Ostensibly because "high school students look up to college students," participants were told that after they viewed the slideshow, they would be asked to write an essay about the possible benefits of the slideshow content for high school students. To help prepare for this essay, intervention participants were given a pen and paper to take notes during the intervention. In the essay, they were allowed to write as little or as much as they wished and were encouraged to cover the major points that were presented during the slideshow. Devine et al. (2012) did not include this essay as a part of the intervention. We added this essay as both a potential measure and because of evidence that self-generated messages can enhance the effect of psychological interventions (Canning & Harackiewicz, 2015; see also Janis & King, 1954; Cialdini, Petty & Cacioppo, 1981). However, adding this essay has the limitation of making this particular instantiation of the habit-breaking intervention less transportable to other contexts. Although we coded the content of the essays, we do not present the coding scheme and results here because there was too little variation in the coding categories for analyses to be meaningful. Nevertheless, the essay coding scheme and results involving the coding categories are presented in full at https://osf.io/a3c8h/.

A link to a follow-up survey was emailed to participants every other evening for two weeks following the in-lab study. Relative to the procedure used by Devine et al. (2012), we shortened the study duration and increased the number and frequency of the follow-ups to obtain a more focused snapshot of the trajectory of change in the study outcomes. We chose a study duration of two weeks and measurement lag of two days to provide a balance between participant burden and measurement frequency. The follow-up survey contained measures of (1) implicit bias, (2) concern about discrimination, (3) discrepancies between standards for and beliefs about one's interracial behavior, and (4) frequencies and content of race-related thoughts, race-related conversations, and interracial interactions involving Black people. Participants assigned to the intervention group also answered questions about each of the five strategies taught during the intervention. They were asked to report how often they used each of the strategies over the past two days using an 8-point scale ranging from "0 times" to "7 or more times", and to briefly describe a situation in which they used the strategy. If participants missed a survey, they were still encouraged to complete the remaining surveys. Across conditions, participants completed an average of 4.79 (SD = 1.75) of the 7 follow-ups. There was no evidence of a differential response rate across conditions (intervention M = 4.90, SD = 1.64, control M = 4.70, SD = 1.84, M_{diff} = 0.20, 95% CI = [−0.21, 0.60].

1.1.3. Implicit Association Test and feedback

We measured implicit bias using the Black-White evaluative IAT (Greenwald et al., 1998). The IAT is a dual categorization task in which people categorize sequentially presented stimuli. In the Black-White evaluative IAT, the participants categorize pictures of Black and White people and pleasant and unpleasant words as to their race and valence, respectively. The assumption underlying the IAT is that people should perform the task faster when concepts that are associated in memory (i.e., Black people with negative words and White people with positive words) are paired on the same response key (*compatible trials*) than with the reverse pairings (*incompatible trials*). Responses on compatible and incompatible trials are used to compute D-scores (Greenwald, Nosek, & Banaji, 2003), which are scored such that higher numbers indicate a greater association of White with positive and Black with negative than the reverse. Overall, participants showed a moderate baseline pro-White bias (baseline M = 0.35, SD = 0.34, skew = −0.50, $r_{\text{split half}}$ = 0.79). Immediately following the IAT, control participants were told their IAT D-scores, along with a short interpretation of the D-score as to whether it indicated a preference for White people, no preference, or a preference for Black people, and whether this preference was strong, moderate, or slight (for more details, see Devine et al., 2012).

1.1.4. Prejudice habit-breaking intervention

Immediately after the IAT, intervention participants completed the prejudice habit-breaking intervention. The habit-breaking intervention is divided into education and training sections. In the education section, the participants learn what implicit bias is, how implicit bias is measured, and the consequences of implicit bias for racial minorities. After

[1] Because participants did not interact with one another in the sessions, there is no a priori reason to believe that there would be interdependence among outcomes at the session level. The random assignment of sessions rather than participants to conditions, however, makes this a cluster-randomized trial. We examined whether there was evidence of within-session interdependence by calculating the session ICC for the IAT, concern, discrepancies, shoulds, and woulds. For the IAT, concern, and woulds, the ICC was effectively 0. For woulds and discrepancies, the ICC was small (woulds ICC = 0.008, discrepancies ICC = 0.069). We therefore treated our design as if randomization occurred at the participant level throughout the paper.

the education section, the participants received feedback about their personal level of implicit bias, which was formatted in a way that was identical to the feedback given to the control participants. Whereas Devine et al. (2012) gave all participants their feedback immediately following the administration of the IAT, we reasoned that the feedback would be more meaningful to participants after they learned about the IAT and what it is purported to measure.

In the training section of the intervention, the participants were introduced to the idea that implicit bias can be overcome through a combination of motivation, awareness, and the use of bias-reduction strategies, and they are taught five such strategies. These strategies were identical to the strategies used by Devine et al. (2012), which included stereotype replacement (Monteith, 1993), counter-stereotypic imaging (Blair, Ma, & Lenton, 2001), individuating (Brewer, 1988; Fiske & Neuberg, 1990), perspective taking (Galinsky & Moskowitz, 2000), and increasing opportunities for contact (Pettigrew, 1998; Pettigrew & Tropp, 2006). To enhance the perceived utility of these strategies, immediately after learning about each one, participants were asked to generate examples of how they could use the strategy in their own life (Hulleman & Harackiewicz, 2009). Participants were informed that the strategies could have synergistic effects and that the more they practiced the strategies, the more effective they would be.

1.1.5. Racial attitudes, IMS, and EMS

These measures were only assessed in the large online survey administered prior to the habit-breaking intervention. Racial attitudes were assessed using a feelings thermometer, which asks participants to rate how warmly they feel toward Black people using a 0 (*very cold*) to 100 (*very warm*) scale. The internal and external motivations to respond without prejudice (IMS and EMS; Plant & Devine, 1998) measure the extent to which people respond without prejudice for internal, value-driven reasons or external, normative reasons. They are assessed with five items each using a 1 (*strongly disagree*) to 9 (*strongly agree*) scale. The measures are scored such that higher numbers indicate more positive attitudes ($M = 76.33$, $SD = 20.27$, skew $= -0.84$), internal motivation ($M = 7.31$, $SD = 1.60$, skew $= -0.82$, $\alpha = 0.76$), and external motivation ($M = 4.17$, $SD = 1.97$, skew $= 0.11$, $\alpha = 0.81$).

1.1.6. Concern about discrimination

Concern measures the extent to which a person believes discrimination toward Black people is a serious problem in society (Devine et al., 2012). The concern scale is composed of 4 items, each of which asks participants to respond to statements like "I consider racial discrimination to be a serious social problem" using a 1 (*strongly disagree*) to 10 (*strongly agree*) scale. Items are scored such that higher numbers indicate greater concern (baseline $M = 6.77$, $SD = 1.96$, skew $= -0.09$, $\alpha = 0.76$).

1.1.7. Discrepancies between standards and beliefs

The discrepancy scale measures the difference between how people believe they *should* act, think, and feel in a variety of race-related situations vs. how they actually *would* act, think, and feel in those same situations (Devine et al., 1991; Monteith & Voils, 1998). The discrepancy scale is divided into separate should and would subscales. In the discrepancy scale developed by Monteith and Voils (1998), each subscale contained 16 items, each of which was measured using a 1 (*strongly disagree*) to 7 (*strongly agree*) scale. We reduced this number to 6 items on the should and would subscales to reduce the scale's total length. The should index is created by averaging the responses on the six should items such that higher scores indicate a greater belief that one should act with bias toward Black people (baseline $M = 1.65$, $SD = 1.02$, skew $= 2.13$, $\alpha = 0.88$). The would index is created by averaging responses on the six *would* items such that higher scores indicate a greater belief that one would act with bias toward Black people (baseline $M = 2.45$, $SD = 1.15$, skew $= 0.76$, $\alpha = 0.84$). The discrepancies index is created by subtracting responses on the should index from responses on the would index. Higher scores indicate a greater belief that one would act with more bias than one believes is appropriate (baseline $M = 0.80$, $SD = 1.18$, skew $= -0.40$, $\alpha = 0.76$).

1.1.8. Race-related thoughts, race-related conversations, and interracial interactions

In the follow-up questionnaires (but not at baseline), participants were asked to report, since the last follow-up questionnaire, the number of thoughts they had had about Black people or issues related to Black people, the number of conversations with others they had had about Black people or issues related to Black people, and the number of interactions they had had with Black people. The length of time "since the last survey" was variable; although the study was designed to obtain follow-up measurements every other day, not all participants completed every survey. The average latency between surveys was slightly longer than 2 days ($M = 2.25$ days, $SD = 1.20$ days). All questions used an 8-point scale ranging from 0 times to 7 or more times.

After each question, if the participants reported at least one thought, conversation, or interaction, they were asked to describe the situation in one to two sentences. We coded these responses for the content of the race-related thoughts, race-related conversations, and interracial interactions by reading through them to identify common themes, which we used to develop a coding scheme. Two people coded each response, and discrepancies between coders were resolved through discussion. We revised the coding criteria of any variables with interrater reliabilities below 0.70 and two new people coded these responses. The final coding scheme and interrater reliabilities are shown in Table 2.[2] Due to high overlap in the thought and conversation themes, we combined scores on these variables for analysis.

2. Results

2.1. Data analytic plan

The intervention and control groups did not differ on any of the variables at baseline (see Table 3), so it appears that random assignment was successful. We conducted all analyses using mixed effects models using lme4 (Bates, Mächler, Bolker & Walker, 2015) and made all plots using ggplot2 (Wickham, 2009). We used Linear Mixed Effects Models if the outcome was quantitative and Generalized Linear Mixed Effects Models if the outcome was not, and obtained confidence intervals using likelihood profiles. Our use of mixed effects has the advantage of using all available information from each participant, thereby providing a natural framework for handling missing data as long they are either missing completely at random or missing at random (MCAR or MAR; Ibrahim & Molenberghs, 2009). For count outcomes, we used a log link in the Poisson family and included an offset for the number of days since the last measurement to account for variable spacing between measurements. For dichotomous variables, we used a logit link in the binomial family.

The random effects structure always included a random intercept for each participant. We included random slopes for predictors that varied within participants. When models with random slopes did not converge, we removed the correlation between the intercepts and slopes, which resulted in convergence in all cases. We note in each section when we excluded the correlations between intercepts and slopes.

[2] Five variables were coded but excluded from analyses. Strategy use was excluded because of low interrater reliability ($\kappa = 0.56$). Thoughts and conversations about the study, whether the race-related conversations occurred in a structured environment, whether the participants confronted another person about their bias, and whether the participants had a negative interracial interaction were excluded because they occurred too infrequently for the GLMEM models to converge (*study* = 60/515, *structured* = 27/249, *confrontation* = 2/249, *negative* = 4/834).

Table 2
Free response coding scheme for the short responses to the race-related thought, race-related conversation, and interracial interaction questions.

Free response item	Variable	Levels	Example	κ
Race-related thoughts; conversations	**Participant bias:** coder decides that participant acted with bias	no = 87% yes = 13%	I was thinking of a joke to say about Black people. [*yes*]	0.74; 0.66
	Other bias: coder decides that a person other than the participant acted with bias	no = 91% yes = 9%	One of my good friends was giving examples of confirmations of African American stereotypes he sees everyday. [*yes*]	0.79; 0.82
	Societal pattern: response notes racial pattern in society	no = 82% yes = 18%	I was studying about rights of African American women in the 1960s. [*yes*]	0.66; 0.66
	Current event: response, mentions current event related to Black people	no = 70% yes = 30%	Talking about Obama being the first Black president [*yes*]	0.79; 0.77
	Labeling: participants labels something as biased/wrong	no = 81% yes = 19%	I was thinking about this survey and thought that I do some racist things [*yes*]	0.79; 0.79
Interracial interactions	**Structured:** interaction occurs in structured context (e.g., classroom, work)	no = 68% yes = 32%	I talked to a Black person in my class yesterday. [*yes*]	0.81
	Unfamiliar: Black interaction partner is unfamiliar (not a friend or family)	no = 80% yes = 20%	I talked to African American customers at work. [*yes*]	0.78
	Quality: interaction is high quality and/or non-superficial	low = 66% high = 34%	I met many of my boyfriend's friends during a visit to West Point and many of them are Black. They were all really nice. And funny. And loud. I liked them! [*high*]	0.65

Table 3
Characteristics of the intervention and control groups at baseline.

	Control				Intervention				Difference
	N	*Mean*	*SD*	*Skew*	*N*	*Mean*	*SD*	*Skew*	
IAT	156	0.37	0.34	-0.65	136	0.34	0.34	-0.32	-.03 [-.11, .04]
Concern	152	6.80	1.85	0.04	130	6.73	2.09	-0.19	-.07 [-.53, .40]
Discrepancies	149	0.88	1.19	-1.01	124	0.71	1.16	0.38	-.17 [-.45, .11]
Shoulds	152	1.58	0.99	2.52	130	1.74	1.06	1.74	.17 [-.07, .41]
Woulds	149	2.46	1.18	0.67	124	2.45	1.14	0.87	-.01 [-.29, .26]
IMS	152	7.36	1.67	-0.99	130	7.25	1.51	-0.56	-.11 [-.49, .27]
EMS	152	4.04	1.95	0.18	130	4.34	1.99	0.02	.30 [-.16, .76]
Attitudes	152	76.91	18.62	-0.60	130	75.65	22.09	-0.97	-1.26 [-6.03, 3.51]

Note: The "Difference" column represents the mean difference in the outcome and its 95% CI.

Unless otherwise noted, all models included as predictors the linear and quadratic effects of time, as well as indicators for condition and the condition by time interactions. When time was included as a predictor, the model also always contained random slopes for time. Time was scaled such that each unit represents one day. Thus, in all analyses of the overall effects of the intervention, the linear effect of time represents the degree to which the outcome of interest changed per day, the quadratic effect of time represents the degree to which the rate of change over time was accelerating or decelerating per day, and the interactions between the time contrasts and condition represents the difference between the intervention and control groups in their rate of change and acceleration/deceleration per day.

Our analyses in Phase 1 address three issues: (1) replication of Devine et al.'s (2012) key findings; (2) examination of people's race-related thoughts, race-related conversations, and interracial interactions; and (3) examination of strategy use.

2.2. Replication analyses

Our first priority was to assess whether we replicated the original test of this intervention. To that end, we examined the primary outcomes measured by Devine et al. (2012), which included implicit bias,

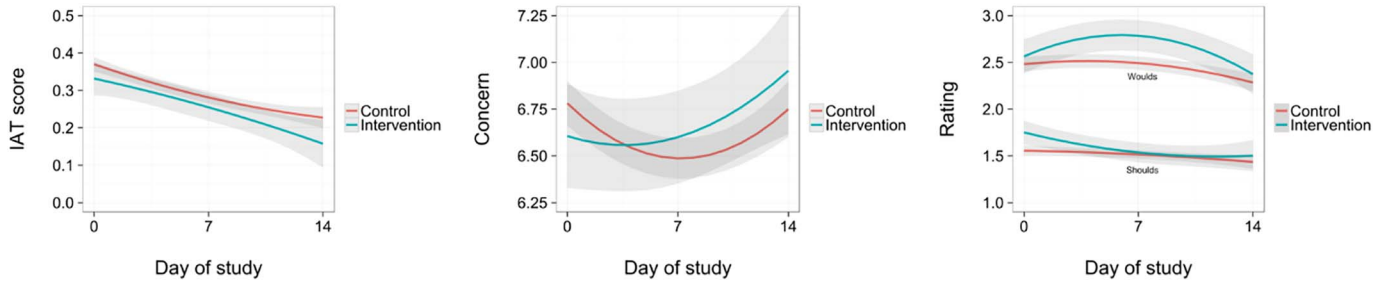

Fig. 2. Changes over time in the intervention and control conditions in IAT scores, concern, and discrepancies (and the components of discrepancies, shoulds and woulds). Envelopes indicate ± 1 Wald standard error of the estimate.

Table 4
Change over time in the IAT, concern, discrepancies, and the components of discrepancies, shoulds and woulds.

Outcome	Condition	Time	Time squared	Condition by time	Condition by time squared
IAT	-.026 [-.093, .041]	-.011 [-.015, -.008]*	.000 [-.001, .001]	-.001 [-.008, .006]	-.001 [-.002, .001]
Concern	.096 [-.351, .542]	.000 [-.015, .015]	.005 [.001, .008]*	.030 [.000, .060]*	-.002 [-.009, .005]
Discrepancies	.267 [.001, .534]*	.009 [-.002, .021]	-.005 [-.008, -.003]*	.022 [.000, .045]	-.007 [-.012, -.002]*
Shoulds	.028 [-.158, .215]	-.015 [-.024, -.006]*	.001 [-.001, .003]	-.012 [-.031, .006]	.002 [-.001, .006]
Woulds	.291 [-.013, .594]	-.006 [-.016, .004]	-.004 [-.006, -.002]*	.009 [-.011, .029]	-.004 [-.008, .000]*

Note: Estimates and their profile likelihood 95% CIs were derived from LMEMs containing condition, linear time, quadratic time, and the interactions between condition and linear/quadratic time. All models contained a random intercept for each participant and random slopes for time and quadratic time.

concern about discrimination, shoulds, woulds, and discrepancies. These model results are shown in Fig. 2 and Table 4.

For many, the most exciting finding from Devine et al. (2012) was that intervention participants, but not control participants, showed a sustained decrease in IAT bias over time. We observed a different pattern in the present study. *Both* intervention and control participants decreased an average of 0.011 IAT units per day, 95% CI = [−0.015, −0.008] – this pattern did not differ for intervention and control participants, $b = -0.001$, 95% CI = [−0.008, 0.006].

Although implicit bias has garnered a tremendous amount of attention, Devine and colleagues have long argued that intentional, conscious processes are essential for breaking the habit of unintentional bias (Devine, 1989; Devine et al., 1991; Devine et al., 2012; Monteith, 1993; Plant & Devine, 1998, 2009). Our measure of concern about discrimination is one indicator of one's conscious belief that discrimination is a serious problem. In the original article, concern significantly increased among intervention but not control participants. We observed the same pattern in the present study: linear change in concern was stronger among intervention than control participants, $b = 0.030$, 95% CI = [0.000, 0.060]. Whereas intervention participants increased in concern over time, $b = 0.015$, 95% CI = [−0.007, 0.036], control participants decreased, $b = -0.015$, 95% CI = [−0.036, 0.005]. Our analyses also revealed a quadratic trend such that people tended to decrease, then increase in concern over time, $b = 0.0056$, 95% CI = [0.0012, 0.0080]. This quadratic trend was not different among the intervention and control participants, $b = -0.0020$, 95% CI = [−0.0088, 0.0049].

The difference between people's reported shoulds and woulds provides an index of whether people believe they would act with more bias than their standards permit in interracial interactions. Devine et al. (2012) found that should–would discrepancies increased for intervention participants but not control participants, an increase that was driven by change in woulds. As shown in Fig. 2, in the present study, we found that although intervention participants initially increased in discrepancies, their discrepancies declined back to pre-intervention levels in the latter parts of the study. This pattern is revealed by the difference in the quadratic trends among intervention and control participants, $b = -0.007$, 95% CI = [−0.012, −0.002]; whereas intervention participants showed a quadratic trend of increases followed by decreases in discrepancies, $b = -0.009$, 95% CI = [−0.012, −0.005], control participants did not, $b = -0.002$, 95% CI = [−0.005, 0.002]. This difference in the quadratic trends for discrepancies was driven by a difference in the quadratic trends for woulds; intervention and control participants showed the same difference in the quadratic trends for woulds that they did for discrepancies, $b = -0.004$, 95% CI = [−0.008, 0.000].

2.3. Replication analyses – discussion

In sum, our findings only partially replicate the findings of Devine et al. (2012). Although intervention participants increased in concern more than control participants, they did not decrease in implicit bias more than control participants. Although intervention participants showed an initial increase in discrepancies relative to control participants, this increase faded in the latter part of the study.

The reason for the inconsistencies between our results and those reported by Devine et al. (2012) is unclear. On the one hand, our sample is more than three times that used by Devine et al. (2012), suggesting that our results are less susceptible to sampling error. The persistent decrease in implicit bias and increase in discrepancies results reported by Devine and colleagues could therefore be false positives. On the other hand, as outlined in Table 1, the current study differed from that conducted by Devine et al. (2012) in six ways, and these differences in procedure could have interfered with the detection of the intervention's effects. In particular, the frequent follow-up assessment schedule could have induced practice effects on the IAT, which may have obscured any true effects of the intervention on implicit bias. Frequently responding to questions about race may have made all participants more sensitive to racial issues, a sensitivity that could have differentially impacted intervention and control participants, perhaps causing the decrease in discrepancies in the latter parts of the study.

Although the interpretation of our findings on implicit bias and discrepancies is unclear, the interpretation of our findings for concern is more straightforward. The habit-breaking intervention appears to have a robust, enduring impact on the degree to which people characterize racial discrimination as a problem. Moreover, some of our other results, reported below, suggest that concern may be as important or more important than implicit bias with regard to empowering people to take steps to address bias. We will return to these issues in the General Discussion.

2.4. Race-related thoughts, race-related conversations, and interracial interactions

We measured race-related thoughts, race-related conversations, and interracial interactions as potential mechanisms for the intervention's effects. We measured both the quantity of these outcomes reported by the participants and the content of these outcomes, as coded from the participants' open-ended descriptions. We first assessed the effects of the habit-breaking intervention on the quantity and content of thoughts, conversations, and interactions, after which we assessed the relationships between the quantity and content variables and implicit bias, concern, and should-would discrepancies.

2.4.1. The habit-breaking intervention's effects on quantity and content

Participants reported a relatively small total number of race-related thoughts ($M = 3.64$, $SD = 5.69$), race-related conversations ($M = 1.86$, $SD = 4.00$), and interracial interactions ($M = 8.13$, $SD = 8.70$) over the course of the study. As shown in Fig. 3, there were no differences across the intervention and control conditions in the change in the daily reported rate of any of the three variables.[3] The only effects that we observed were steady decreases in the daily reported rate of thoughts $b = 0.871$, 95% CI = [0.843, 0.899], conversations

[3] We did not estimate the correlations between random slopes and random intercepts in these models.

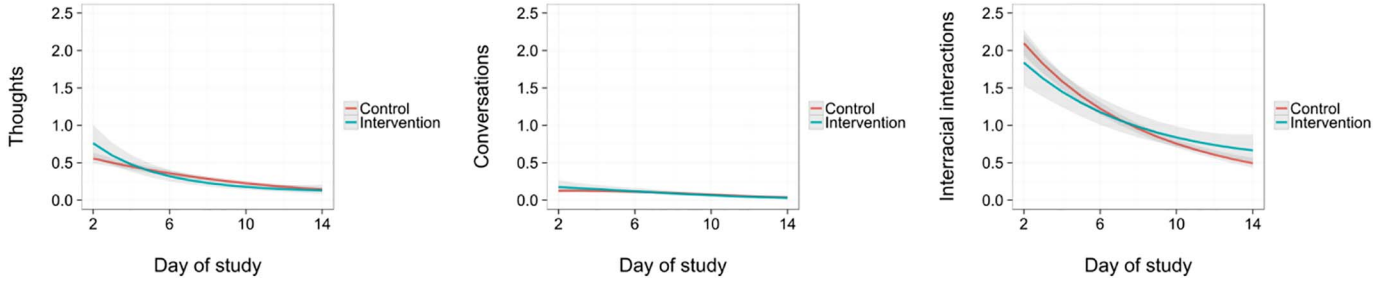

Fig. 3. Changes over time in the control and training conditions in the number of reported daily race-related thoughts, race-related conversations, and interracial interactions. Envelopes indicate ± 1 Wald standard error of the estimate.

Table 5
The relationships between the quantity and content of race-related thoughts, conversations, and interactions and the IAT, concern, discrepancies, and the components of discrepancies, shoulds and woulds.

	Source	Predictor	IAT	Concern	Discrepancies	Shoulds	Woulds
Quantity	Thoughts	Rate per day (average)	.001 [-.020, .034]	-.073 [-.157, .020]	-.015 [-.078, .054]	.015 [-.037, .067]	-.001 [-.062, .046]
		Rate per day (change)	-.019 [-.072, .034]	.261 [-.118, .639]	.114 [-.109, .337]	.001 [-.020, .034]	.001 [-.247, .265]
	Conversations	Rate per day (average)	.033 [-.017, .085]	-.078 [-.253, .104]	-.088 [-.174, -.001]*	.094 [-.004, .186]	.009 [-.064, .084]
		Rate per day (change)	-.013 [-.087, .060]	.026 [-.509, .561]	.117 [-.196, .430]	.011 [-.231, .209]	.100 [-.260, .460]
	Interactions	Rate per day (average)	-.019 [-.041, .002]	-.004 [-.064, .053]	-.053 [-.097, -.011]*	-.002 [-.037, .032]	-.059 [-.099, -.020]*
		Rate per day (change)	-.044 [-.078, -.011]*	.167 [-.081, .416]	-.055 [-.201, .090]	-.106 [-.207, -.004]*	-.164 [-.330, .003]
Content	Thoughts/conversations	Participant bias	.049 [-.078, .177]	-1.716 [-2.668, -.762]*	.432 [-.137, .999]	.338 [-.052, .727]	.761 [.118, 1.403]*
		Others' bias	-.004 [-.143, .135]	1.476 [.421, 2.531]*	.610 [-.009, 1.228]	-.298 [-.725, .128]	.296 [-.417, 1.009]
		Societal pattern	-.060 [-.163, .043]	.907 [.117, 1.696]*	.228 [-.236, .692]	-.327 [-.643, -.011]*	-.095 [-.625, .435]
		Current event	.000 [-.088, .089]	.273 [-.407, .955]	-.126 [-.522, .271]	-.244 [-.515, .026]	-.367 [-.817, .084]
		Labeling	-.035 [-.141, .071]	1.650 [.862, 2.437]*	.040 [-.436, .517]	-.405 [-.727, -.083]*	-.359 [-.901, .184]
	Interactions	Structured context	-.006 [-.085, .073]	.170 [-.428, .769]	.004 [-.335, .341]	-.120 [-.339, .100]	-.111 [-.498, .275]
		Unfamiliar	.104 [-.011, .196]	-.645 [-1.340, .052]	.331 [-.064, .726]	.004 [-.254, .262]	.347 [-.105, .798]
		High quality	-.069 [-.152, .014]	.556 [-.068, 1.181]	-.305 [-.658, .048]	-.132 [-.363, .099]	-.441 [-.844, .038]

Note: The quantity variables were calculated by dividing the reported quantities of race-related thoughts and conversations and interracial interactions and dividing by the elapsed time since the last measurement of these variables. We used each participant's average rate of change and the differences between their average rate and their occasion-specific rates as predictors in an LMEM, with random slopes for the centered occasion-specific rates. For the content variables, we calculated the proportion of each participant's descriptions that were coded in each category and used these proportions as predictors in an LMEM. All 95% CIs were derived using profile likelihood. All models contained effects for condition, linear and quadratic effects of time, interactions between condition and linear and quadratic time, random intercepts for each participant, and random slopes for linear and quadratic time.

b = 0.890, 95% CI = [0.845, 0.931], and interactions b = 0.895, 95% CI = [0.877, 0.911]. Full results for these models are available at https://osf.io/a3c8h/.

Although the habit-breaking intervention did not affect the *number* of thoughts, conversations, and interactions that occurred over the study's two-week duration, it may have affected what happened during these incidents. As a reminder, participants only described their thoughts, conversations, and interactions when they reported that thoughts, conversations, or interactions had occurred since their last survey. People reported a non-zero number of thoughts, conversations, and interactions a relatively small number of times, resulting in a small number of codeable descriptions per person (thoughts M = 1.60, conversations M = 0.90, interactions M = 2.86). We therefore did not investigate time trends in these analyses.

Intervention participants were more likely to mention an incident in which a coder identified that someone other than the participant acted with bias, $p_{control}$ = 0.032, $p_{intervention}$ = 0.082,[4] OR = 2.710, 95% CI = [1.227, 6.626]. They were also more likely to label biases, whether committed by themselves, someone else, or observed in society, as wrong, $p_{control}$ = 0.077, $p_{intervention}$ = 0.171, OR = 2.468, 95% CI = [1.278, 4.984]. Finally, intervention participants were more likely to mention that their interracial interactions were with relative strangers, $p_{control}$ = 0.058, $p_{intervention}$ = 0.021, OR = 0.340, 95% CI = [0.117, 0.913], though the absolute difference in predicted probabilities was small. All other content comparisons between control and intervention participants were non-significant and are shown in full at https://osf.io/a3c8h/.

2.4.2. Relationships with the main study outcomes

A predictor measured multiple times for each person can be associated with an outcome in two ways (Raudenbush & Bryk, 2002). First, a person's *average level* of the predictor – in other words, the variance between people – can be associated with the outcome. Second, a person's *change* in the predictor – in other words, the variance within people – can be associated with the outcome. To examine whether these components of the quantity of race-related thoughts, race-related conversations, and interracial interactions were associated with the IAT, concern, should, woulds, and discrepancies, we first put the quantity variables on a common metric by dividing them by the amount of time that had passed since the last measurement, resulting in rates per day.[5] We then constructed indicators of between-person variance by finding, for each person, their average daily rates of thoughts, conversations, and interactions. We constructed indicators of within-person change by, for each person, subtracting their mean rates from each of their occasion-specific rates (Enders & Tofighi, 2007). Finally, we fit separate models, each of which simultaneously predicted each of our outcome

[4] These numbers represent predicted values from the GLMEM and therefore cannot be interpreted directly as percentages.

[5] Some participants took multiple follow-ups in the same day, which resulted in erroneously high daily rates due to the short amount of elapsed time between follow-ups. Thus, we excluded responses that occurred less than half a day after the last recorded follow-up from these analyses. We did not estimate the correlations between the random intercepts and slopes in these models.

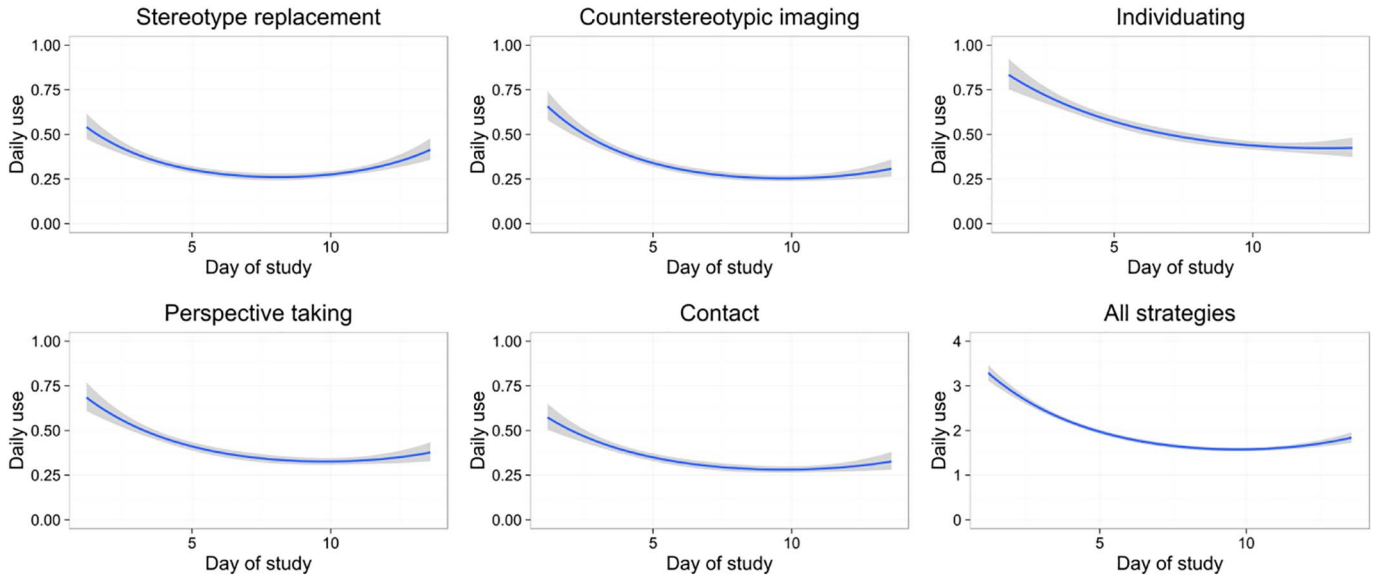

Fig. 4. Changes over time in the daily usage of each strategy. Envelopes indicate ± 1 Wald standard error of the estimate.

variables from the between-person and within-person indicators for either thoughts, conversations, or interactions.

As shown in Table 5, controlling for condition and its interactions with time, the quantity of race-related thoughts and conversations was not associated with any of the main study outcomes. The one exception was that a person's average daily rate of race-related conversations was associated with smaller should-would discrepancies, $b = -0.088$, 95% CI = [−0.174, −0.001]. In contrast, the quantity of interracial interactions was more strongly associated with the main study outcomes. Controlling for condition and its interactions with time, a person's average level of interracial interactions was associated with smaller should-would discrepancies, $b = -0.053$, 95% CI = [−0.097, −0.011], a relationship that was driven by lower levels of woulds, $b = -0.059$, 95% CI = [−0.099, −0.020]. Within-person increases in the daily rate of interactions was associated with both change in IAT bias, $b = -0.044$, 95% CI = [−0.078, −0.011], and increasingly strict standards for one's race-related behavior, $b = -0.106$, 95% CI = [−0.207, −0.004]. None of the quantity variables was associated with concern.

We also assessed whether the content of the participants' race-related thoughts, race-related conversations, and interracial interactions was associated with the main study outcomes. As a reminder, we could only assess content when the participants reported a non-zero number of thoughts, conversations, and interactions. This means that we only had a sufficient number of responses for each person to estimate average content for each person, rather than both the average content and the within-person change in content. We estimated between-person content by calculating the proportion of each participant's responses that fell in each coding category.

As shown in Table 5 and in contrast to the quantity variable results, several of the content variables were associated with to follow-up concern about discrimination, even controlling for condition and its interactions with time. A higher proportion of thought and conversation descriptions in which the participants themselves acted biased was associated with lower levels of follow-up concern, $b = -1.72$, 95% CI = [−2.67, −0.76]. Noting the position of Black people in society was associated with higher concern, $b = 0.91$, 95% CI = [0.12, 1.69], as was noticing others act with bias, $b = 1.48$, 95% CI = [0.42, 2.53], and labeling bias as wrong, $b = 1.65$, 95% CI = [0.86, 2.44]. The relationships between noticing others act with bias and labeling bias as wrong on the one hand and concern on the other are particularly interesting given that the habit-breaking changed these outcomes. Due to ambiguity in how to estimate, in a nested design, the indirect effects of dichotomous mediators in the presence of variably spaced time points across participants, we did not conduct a formal mediation analysis. Nevertheless, our combined results suggest that the habit-breaking intervention exerts its impacts on concern because it orients people to the behavior of others and encourages people to label behaviors as explicitly wrong.

In addition to these relationships with concern, the proportion of descriptions in which the participants reported that they acted with bias was positively associated with woulds, $b = 0.76$, 95% CI = [0.12, 1.40], and the proportion of incidents that were labeled as wrong was negatively associated with shoulds, $b = -0.41$, 95% CI = [−0.73, −0.08]. Finally, noticing patterns involving Black people in society was associated with stricter standards for one's race-related behavior, $b = -0.33$, 95% CI = [−0.64, −0.01]. There were no relationships between the content of interracial interactions and the main study outcomes.

2.5. Strategy use

Whereas the analyses above include both intervention and control participants, the remaining set of Phase 1 analyses include only intervention participants. These analyses assess the intervention participants' patterns of strategy use, as well as whether strategy use is associated with the main study outcomes.

2.5.1. Patterns of strategy use

Participants reported using one of the strategies an average of 8.93 times ($SD = 13.51$) during the two weeks following the intervention. Each individual strategy was used infrequently — for all five strategies, the modal amount of strategy use was 0, and all strategies except individuating were used less than two times on average; stereotype replacement $M = 1.50$, $SD = 2.87$, counterstereotypic imaging $M = 1.54$, $SD = 2.80$, individuating $M = 2.46$, $SD = 4.52$, perspective taking $M = 1.85$, $SD = 3.23$, contact $M = 1.58$, $SD = 3.48$.

As shown in Fig. 4, total strategy use was highest at the start of the study and decreased daily by a factor of 0.881, 95% CI = [0.846, 0.916]. This pattern of constantly decreasing usage was repeated across all five strategies, stereotype replacement $b = 0.953$, 95% CI = [0.905, 0.999], counterstereotypic imaging $b = 0.914$, 95% CI = [0.867, 0.960], individuating $b = 0.953$, 95% CI = [0.915, 0.988], perspective taking $b = 0.926$, 95% CI = [0.882, 0.967], contact $b = 0.920$, 95% CI = [0.862, 0.972]. This pattern could either reflect an actual decrease in usage or survey fatigue, given that the survey solicited a description

Table 6
The relationships within the intervention condition between strategy use and the IAT, concern, discrepancies, and the components of discrepancies, shoulds and woulds.

	Predictor	IAT	Concern	Discrepancies	Shoulds	Woulds
Replacement	Average	-.014 [-.181, .152]	-.303 [-1.450, .846]	.284 [-.411, .978]	-.109 [-.614, .395]	.141 [-.626, .899]
	Change	.030 [-.044, .108]	-.026 [-.272, .220]	.283 [.040, .542]*	-.146 [-.298, -.010]*	.048 [-.138, .248]
Imaging	Average	.010 [-.166, .187]	-.944 [-2.159, .273]	.380 [-.360, 1.118]	.050 [-.485, .585]	.398 [-.410, 1.207]
	Change	-.007 [-.085, .101]	-.340 [-.692, -.020]*	.121 [-.068, .345]	.017 [-.127, .193]	.168 [-.050, .405]
Individuating	Average	-.044 [-.135, .047]	-.501 [-1.123, .122]	-.048 [-.430, .334]	-.145 [-.421, .131]	-.181 [-.596, .234]
	Change	.024 [-.018, .070]	.079 [-.100, .268]	-.034 [-.141, .068]	-.015 [-.098, .057]	-.049 [-.180, .038]
Persp. taking	Average	-.081 [-.238, .077]	.279 [-.709, 1.469]	-.131 [-.788, .536]	-.057 [-.531, .416]	-.216 [-.936, .502]
	Change	.052 [-.014, .135]	.146 [-.169, .499]	.016 [-.189, .231]	-.084 [-.209, .032]	-.104 [-.243, .035]
Contact	Average	-.139 [-.294, .017]	.267 [-.822, 1.357]	-.117 [-.774, .539]	-.089 [-.562, .384]	-.190 [-.909, .528]
	Change	.021 [-.072, .121]	-.094 [-.376, .171]	.068 [-.100, .241]	-.033 [-.155, .087]	.035 [-.141, .229]
All strategies	Average	-.019 [-.055, .017]	-.100 [-.352, .152]	.009 [-.143, .162]	-.035 [-.145, .075]	-.028 [-.194, .139]
	Change	.019 [-.008, .049]	-.004 [-.089, .086]	.024 [-.029, .092]	-.021 [-.056, .012]	-.010 [-.051, .036]

Note: Each strategy use variable was derived by dividing the number of uses of the strategy by the time since the last measurement. We then used both each participant's average daily rate and the differences between their occasion-specific daily rates and this average as predictors in an LMEM, along with a random slope for the centered occasion-specific rates. All 95% CIs were derived using profile likelihood. All models also contained effects of condition, linear and quadratic time, a random intercept for each participant, and random slopes for linear and quadratic time.

of where and when a strategy was used only if any usage was reported. There were no curvilinear usage effects. None of the strategies had an average rate of use that was greater than 0.5 per day at any time during the study.

2.5.2. Relationships with the main study outcomes

Just as with the quantities of race-related thoughts, race-related conversations, and interracial interactions, both a person's average rate of strategy use and the within-person change in this rate could be associated with the main study outcomes. We therefore calculated, for each strategy, each person's average daily rate of use, as well as the differences between a person's occasion-specific rates and their average rate (Enders & Tofighi, 2007). We then fit separate models, each of which simultaneously predicted the IAT, concern, shoulds, woulds, and discrepancies from the between-person and within-person indicators for use of one of the five strategies.[6]

As shown in Table 6, change in the rate of strategy use was largely unimportant for predicting change in the IAT, concern, shoulds, woulds, and discrepancies. This was true regardless of whether we examined people's average rate of strategy use or their changes in strategy use, and regardless of whether we examined each strategy individually or the total use of all strategies. There were two exceptions. First, increases in the rate of stereotype replacement usage were related to increases in discrepancies, $b = 0.283$, 95% CI = [0.040, 0.542]. This was driven by a negative association between stereotype replacement usage and shoulds, $b = -0.146$, 95% CI = $[-0.298, -0.010]$. Second, increases in the rate of counterstereotypic exemplar usage were related to *decreases* in concern about discrimination, $b = -0.340$, 95% CI = $[-0.682, -0.020]$. It is possible that these two relationships indicate that adopting stricter standards for one's interracial behavior leads to more frequent stereotype replacement and decreasing concern leads to more frequent counterstereotypic imaging. However, it is also possible that stereotype replacement has a beneficial causal effect on standards, perhaps by making standards salient on a frequent basis, whereas counterstereotypic imaging has an ironic negative causal effect on concern, perhaps by highlighting examples of outgroup members who appear unaffected by discrimination.

[6] As with our other analyses using rate variables, we excluded responses that occurred less than half a day after the previously recorded response.

3. Phase 2

3.1. Overview

Two years after Phase 1, we conducted an exploratory study in which we measured the main study outcomes from Phase 1 and three behaviors that could plausibly be affected by the habit-breaking intervention. To that end, participants from Phase 1, who were kept unaware of the new study's connection to Phase 1, were invited to participate in a survey about student engagement with issues that affect the university. They learned we were evaluating student interest in a potential new online section of the campus student newspaper. As part of this evaluation process, the participants read an essay, supposedly written by another student, that argued that racial stereotypes are harmless, cognitively efficient, and are unpopular only because of a desire to be politically correct. Participants then had a chance to both privately rate their agreement with the essay and publicly post a comment about the essay's content. Finally, at the end of the study, they were given the chance to donate any amount of their compensation to a charity that has the mission of eliminating racial discrimination.

As part of its emphasis on awareness, the slideshow that forms the basis of the habit-breaking intervention specifically mentions the importance of detecting and labeling as wrong instances of bias in the social environment. We therefore reasoned that, to the extent that the habit-breaking intervention has effects that endure, intervention participants should be more likely to privately disagree with the content of the essay. If the habit-breaking intervention also leads people to make their views public, intervention participants may also be more likely to disagree with the essay publicly in their comments. Because reading and responding to written pieces on blogs and social networking websites is a common experience for many, this measure has a close match to people's everyday experiences. Finally, because the habit-breaking intervention increases concern about racial discrimination, we reasoned that intervention participants may donate more to organizations aimed at eliminating this discrimination.

3.2. Recruitment

All 292 participants from Phase 1 were eligible for Phase 2. We sent the participants an initial email to recruit them to take a survey about student engagement with issues that affect the university. The recruitment email mentioned that participants would be paid $10 for their time, and that they would be given the opportunity to donate any

amount of this $10 to charity. The email did not mention anything connected to Phase 1.

After the initial recruitment email, we made strenuous efforts to obtain a high response rate; specifically, we sent two subsequent reminder emails and, if we had a cell phone number for our participants from Phase 1, we made a personal appeal to participate by cell phone. Of the 292 participants from Phase 1, 108 (42 intervention, 66 control, 74% female, 73% White, 19% Asian) consented to Phase 2, yielding a 37% response rate. Of the 108 consenting participants, 77 (71%) completed the survey, and 39 (36%) completed an IAT to which the participants were redirected after the completion of the survey. On average, participants started Phase 2 two years (M = 786 days, SD = 76 days) after they started Phase 1.

Assuming an effect size typical of social psychology (r = 0.21; Richard, Bond, & Stokes-Zoota, 2003), our sample of 77 yields 46% power to detect an effect of the habit-breaking intervention on one of our behavioral measures. If we instead assume an effect size closer to the ones obtained by Devine et al. (2012) on the IAT (r = 0.28), concern (r = 0.22), and discrepancies (r = 0.22), our power ranges from 51% to 71%. Phase 2 is therefore somewhat underpowered, meaning that significance tests in Phase 2 cannot distinguish between effects that are small and effects that are non-existent. Moreover, any effects that we do detect may be overestimates (Button et al., 2013). Despite these limitations, we believe the rarity of long-lag data on the effectiveness of bias interventions make these data worth examining.

3.3. Procedure

All materials and data are publicly available at https://osf.io/a3c8h/. Participants were informed that the study was testing a new online section of the campus student newspaper, called “Dialogue,” that would feature a weekly essay from a student on a topic of importance to the campus community. Students reading the “Dialogue” section would have the chance to discuss the essay in a comments section below the essay.

The participants were then asked to read a sample essay for the dialogue section entitled “Racial stereotypes are useful tools” that argued that stereotypes are useful for bypassing the effortful process of treating all people as individuals. The essay further argued that stereotyping has only become “untrendy” because our society is too politically correct and that stereotypes are harmless.

After reading the essay, participants commented on the essay, reported whether they agreed with the essay, and completed the survey measures from Phase 1. These survey measures included concern, shoulds, and woulds, as well as the measures only measured at baseline, namely IMS, EMS, and attitudes toward Black people. At the end of the survey, participants were asked whether they would like to donate any or all of their earnings to any of 4 potential charities, including a charity that had the goal of reducing racial discrimination. Following the donation measure, participants completed the IAT.

3.3.1. Essay agreement

The participants were asked to rate their agreement with seven statements using a 1 (*strongly disagree*) to 7 (*strongly agree*) scale. They were assured that these ratings would remain confidential. Six of these statements were filler items designed to enhance the cover story (e.g., “I think students will use this website”, “I would be interested in writing essays for this section of the paper”), whereas one of the items was the item of interest (“I agreed with the author's main points”).

3.3.2. Essay comments

The participants were asked to post a comment in response to the essay, which they were told would go live with the essay if and when the “Dialogue” section was added to the student newspaper. Two independent coders categorized each comment according to whether it expressed disagreement with the content of the essay (59.5%), was neutral toward the content of the essay (17.7%) or expressed an opinion that was mixed (22.8%). No commenter expressed unreserved agreement with the essay. A sample disagreeing comment is:

Although I see the point you are trying to make about how stereotyping can be a useful tool, I completely disagree that stereotypes give a good indication of what any particular person will be like. Outward appearances very rarely reflect the true depth of any individual, and often times these stereotypes are built around themes we've seen in the media. Also, the idea that a stereotype is not going to affect how you act toward a person has been proven wrong through many different scientific experiments that have found that stereotypes highly influence our actions toward others.

A sample neutral comment is:

This opinion would get quite a response from the ethnic based student organizations and advocacy groups. It is definitely a way to get people talking, but it could really start a negative reaction toward the Badger Herald.

A sample mixed comment is:

I thought it was ignorant at first but believe the reasoning behind stereotypes is very true, not necessarily saying it's a good thing or bad thing.

We combined the neutral and mixed categories for the purposes of analysis.

3.3.3. Donations

The participants were given the opportunity to donate any amount of their compensation (including $0) to each of four charities. The participants were given short descriptions of the goals of each charity and links to the charity websites. One of the charities, the Center for Social Inclusion, has the goal to eliminate racial discrimination in policies affecting transportation, food, and housing opportunities for people of color. The other three charities, the World Wildlife Fund, the American Red Cross, and the Make a Wish Foundation, were included to enhance the cover story. The dependent variable of interest was the dollar amount of donations to the Center of Social Inclusion.

3.3.4. Phase 1 outcomes

We measured the main outcomes from Phase 1, namely implicit bias, concern, and should-would discrepancies. We also measured the outcomes that were measured at baseline but not the Phase 1 follow-ups, namely racial attitudes and the motivations to respond without prejudice. The IAT was the last measure we administered in the survey. Due to limitations in the survey software, the participants had to be redirected to a separate website, where they were prompted to download a plugin required to administer the IAT. Presumably because taking the IAT involved extra effort, only 39 participants completed it.

3.4. Results

As shown in Table 7, participants who consented to Phase 2 were equivalent in their baseline characteristics. For all variables except comments, we estimated the difference between the intervention and control participants using a General Linear Model. For comments, we used a Generalized Linear Model with a logit link from the binomial family. If the outcome measure in question was measured at baseline in Phase 1, we used the baseline measurement as a covariate to increase power (Van Breukelen, 2006).

3.4.1. Longevity of change in Phase 1 outcomes

As shown in Table 7, there was modest evidence that intervention participants had greater reported woulds than control participants, b = 0.499, 95% CI = [−0.003, 1.001], though the 95% CI for this difference overlapped slightly with 0. There was no evidence of any other differences between the Phase 1 outcomes, though the descriptive difference in concern (intervention M = 7.39, SD = 1.64; control M = 6.72, SD = 2.53) was in the predicted direction. Our somewhat

Table 7
Comparisons between the intervention and control conditions among the consenting participants for Phase 2.

		Control				Intervention				Difference
		N	*Mean*	*SD*	*Skew*	*N*	*Mean*	*SD*	*Skew*	
Baseline	IAT	66	0.33	0.33	-0.17	42	0.33	0.30	-0.04	0.00 [-0.13, 0.12]
	Concern	66	6.62	1.68	0.07	42	6.87	1.83	-0.20	0.25 [-0.43, 0.93]
	Discrepancies	65	0.85	1.10	0.44	41	0.79	1.01	0.19	-0.06 [-0.48, 0.36]
	Shoulds	66	1.59	0.90	1.91	42	1.54	0.90	2.01	-0.05 [-0.40, 0.30]
	Woulds	65	2.45	1.14	0.69	41	2.34	1.01	0.63	-0.11 [-0.54, 0.32]
	IMS	66	7.35	1.62	-0.90	42	7.45	1.40	-0.83	0.10 [-0.50, 0.70]
	EMS	66	3.94	1.88	0.32	42	4.31	2.07	0.09	0.37 [-0.39, 1.14]
	Attitudes	66	75.94	18.24	-0.46	42	75.83	18.90	-0.48	-0.11 [-7.35, 7.13]
Follow-up	IAT	23	0.06	0.51	-0.51	16	0.18	0.42	-0.69	0.11 [-0.20, 0.41]
	Concern	51	6.72	2.53	-0.62	29	7.39	1.64	0.17	0.65 [-0.39, 1.69]
	Discrepancies	50	1.40	1.17	0.65	28	1.85	1.08	0.11	0.44 [-0.08, 0.95]
	Shoulds	51	1.24	0.39	1.93	28	1.20	0.28	1.09	-0.02 [-0.18, 0.14]
	Woulds	50	2.64	1.28	0.59	29	3.03	1.18	0.18	0.50 [0.00, 1.00]
	IMS	48	7.45	1.73	-1.25	26	7.37	1.41	-1.47	-0.14 [-0.88, 0.60]
	EMS	48	4.54	2.49	0.14	26	4.71	2.19	0.39	-0.05 [-1.19, 1.09]
	Donation	52	1.06	2.06	2.28	30	1.10	2.26	2.38	0.04 [-0.93, 1.02]
	Essay agreement	52	2.71	1.86	0.74	30	2.10	1.12	0.80	-0.61 [-1.36, 0.13]
	Disagreeing comments	50	0.48	--	--	29	0.79	--	--	4.15 [1.51, 12.84]*

Note: For the baseline variables, the "Difference" column shows the mean difference between the intervention and control participants and its 95% CI. For the follow-up variables, the "Difference" column shows the difference and 95% CI, controlling for the baseline measurement of the variable in question (if the variable was measured at baseline). For the disagreeing comments variable, this difference is an odds ratio rather than a mean difference.

low power means that it is ambiguous whether the difference in concern that we observed in Phase 1 has faded by Phase 2 or if our sample is simply too small to detect an enduring difference.

3.4.2. Behavioral measures

Both intervention and control participants disagreed privately with the essay's content; rated agreement in both groups was well below the scale midpoint of 4 (M = 2.48, SD = 1.65). There was no evidence that the degree of private disagreement differed by condition, intervention M = 2.10, SD = 1.12, control M = 2.71, SD = 1.86, M_{diff} = 0.61, 95% CI = [−0.13, 1.36]. However, there was evidence that intervention participants were more likely to post a public comment disagreeing with the premise that stereotypes were harmless. Whereas 48% of control participants wrote a disagreeing comment, 79% of intervention participants wrote one, OR = 4.15, 95% CI = [1.51, 12.84]. Only 25 participants chose to donate any amount to the Center for Social Inclusion, and the amount donated did not differ by condition, intervention M = $1.10, SD = $2.26, control M = $1.06, SD = $2.06, M_{diff} = $0.04, 95% CI = [$-0.93, $1.02].

4. General discussion

Consistent with the results reported by Devine et al. (2012), we found that the habit-breaking intervention produced an enduring impact on concern about discrimination. Although the present study replicated the intervention's effect on concern, its results related to should-would discrepancies and the IAT were not fully consistent with those reported by Devine and colleagues. Devine and colleagues found that the habit-breaking intervention produced an enduring increase in discrepancies by increasing woulds, whereas we found that the initial increase in woulds declined back to baseline in the latter part of the study. In addition, Devine and colleagues found that only intervention participants declined in their IAT scores, whereas we found that both intervention and control participants exhibited this decrease.

Extending the original work, we found that in the two weeks following the manipulation, intervention participants were more likely to (1) notice bias in the world around them, (2) label any bias (in themselves, others, or society) as wrong, and (3) have interracial interactions with relative strangers (as opposed to friends and family). Two years later, intervention participants were more likely to confront bias by writing comments disagreeing with an essay advocating stereotyping. We believe that, despite some ambiguities in our findings, they provide compelling evidence that the prejudice habit-breaking intervention causes its recipients to recognize bias and its consequences for minorities, then address it in the world around them.

4.1. Replication inconsistencies

The present study has a much larger sample size than the study reported by Devine and colleagues. This means that, as long as the present study is a fair test of the effects of the habit-breaking intervention, it is more likely that the effects on implicit bias and discrepancies reported by Devine and colleagues are false positives than that the present results are false negatives. However, the six differences in procedure outlined in Table 1 complicate this interpretation – the new procedural elements in the present study may have masked the habit-breaking intervention's true effects. Of these differences, we believe those most likely to interfere are the changes to the follow-up administration schedule. For example, the frequent assessment schedule may have focused the participants' attention on race-related issues, which could have interfered with the effect of the intervention on should-would discrepancies. After the initial spike in discrepancies, intervention participants may have either worked to reduce their biases or compared their own behavior to the new biases that they observed in other people, either of which may have caused them to revise their beliefs about their biases back to their baseline levels. As another example, the high frequency of IAT administration could have given participants sufficient practice to "beat" the IAT (Banse, Seise, & Zerbes, 2001; Kim, 2003; Lai et al., 2014; Steffens, 2004). If these practice effects are not additive with the habit-breaking intervention's true effects, they could result in similar decreases in IAT scores in both the control and the intervention conditions, which would mask a true non-zero effect of the intervention on implicit bias.

Ultimately, the precise reasons for the differences between our findings and those reported by Devine et al. (2012) are ambiguous. Regardless, we replicated the original effect of the habit-breaking

intervention on concern, which gives us confidence that this effect is robust. The habit-breaking intervention's impact on concern amounted to an estimated difference of 0.42 scale units at the end of the 14-day period of Phase 1. Until we obtain population-level evidence about the typical distribution of concern in representative samples, the pragmatic importance of a difference this size is unclear. However, as we will describe in more detail below, both our mechanism analyses and general theory about the self suggest that producing change in concern may have considerable theoretical importance to the bias reducing process. As we will argue, this theoretical importance may eclipse that of implicit bias.

4.2. The habit-breaking intervention's mechanisms of change

Our analyses of reported strategy use and the quantity of race-related thoughts, race-related conversations, and interracial interactions suggest that none of these is likely responsible for producing the intervention's effects. The habit-breaking intervention does not change these variables, and with the exception of the quantity of interracial interactions, few of these variables were associated with implicit bias, concern, or discrepancies. In contrast, three dimensions of the content of race-related thoughts, race-related conversations, and interracial interactions differed as a function of the intervention. Participants who completed the habit-breaking intervention were slightly more likely to mention spending time with unfamiliar Black people. Intervention participants were also more likely to mention instances in which others acted with bias and more likely to label any biases they observed (in themselves, others, or society) as wrong.

In particular, noticing the biases of others and labeling biases as wrong were both associated with follow-up concern, even controlling for the intervention's effects. These results suggest that noticing the biases of others and labeling biases as wrong serve as mechanisms through which the habit-breaking intervention increases the degree to which people think discrimination is a problem. If increased concern provokes people to pay more attention to biases in the external environment, concern could be implicated in a recursive, self-sustaining process. A recursive relationship between concern and sensitivity to bias in the external environment could be one reason for the persistence of the intervention's effects. If true, these speculations suggest a potential means through which researchers and practitioners could further enhance the intervention's effectiveness: encourage people to attend to the biases of others in their daily life.

Overall, our results suggest that the habit-breaking intervention does not primarily exert its effects through the strategies the participants learn, nor does it exert its effects by changing the raw quantity of race-related thoughts, race-related conversations, or interracial interactions. Instead, the habit-breaking intervention increases people's sensitivity to bias, particularly when others act with bias, and increases the probability that, when a person encounters bias, he or she will label that bias as wrong. The process of detecting bias in others and labeling it as wrong may in turn provoke concern about racial discrimination, just as the concern may itself provoke the detection of future bias.

4.3. The habit-breaking intervention and long-term change

Finally, we uncovered evidence that some of the impact of the habit-breaking intervention is truly long-lasting. Intervention participants were more likely than control participants to publicly object to an essay arguing that stereotypes are useful two years after the administration of the intervention. It is noteworthy that the process of identifying the biased behavior in another person, labeling it as such, and speaking out about it in a public forum is quite similar to the processes that we uncovered in the content of participants' race-related thoughts and conversations from Phase 1. In this sense, the commenting measure provides a good match with the processes that underlie the habit-breaking intervention effects, and the difference in objecting comments at Phase 2 suggests that these processes persisted two years after the intervention was administered. However, we temper this interpretation by noting that our Phase 2 sample was somewhat small, and the true effect of the habit-breaking intervention on the tendency to object to biases may therefore be somewhat smaller than we report here (Button et al., 2013). Nevertheless, we believe the Phase 2 data provide tentative, encouraging evidence as to the longevity of the habit-breaking intervention's effects.

Why do the effects of the habit-breaking intervention persist? One compelling account stems from Rokeach's theory of the self (Rokeach, 1973). This theory holds that behavior is governed by a self-system that is arranged hierarchically around a person's self-concept (Rokeach, 1973). All aspects of the self are affected to some degree by the social environment. However, the aspects that are most central to the self-concept, such as values, are highly resistant to environmentally induced change precisely because such changes require the reorganization of other components lower in the hierarchy. By the same token, the more peripheral aspects of the self, such as stereotypic associations, are highly susceptible to environmental influence because they require no such reorganization (Forscher et al., 2017).[7]

Rokeach's theory offers a useful lens for thinking about psychological change processes. From the perspective of a person designing an intervention, the decision as to whether to target processes central or peripheral to the self-concept offers a tradeoff. Centrally located processes (e.g., values) may be more difficult to change initially, but once initiated, a change can spur a large-scale reorganization of the self-system (Rokeach, 1973). Supported by the newly restructured self-system, the change is also likely to endure despite possibly countervailing environmental influences. In contrast, peripheral processes (e.g., stereotypic associations) are highly susceptible to intervention, at least in the short term. However, because these changes are not buffered by a newly reorganized self-system, they are more likely to be erased by countervailing environmental influences with the passage of time (Lai et al., 2016).

If we apply this analysis to the habit-breaking intervention, we see that the primary targets of change are not the processes that are central to the self-concept. Indeed, in both the data presented by Devine et al. (2012) and in our Phase 2 data, the internal motivation to respond without prejudice, which stems from a person's core egalitarian values, did not change as a function of the intervention. Likewise, the habit-breaking intervention does not directly target automatic processes like implicit bias. Rather, the primary targets of change are beliefs or knowledge – knowledge about how biases can affect behavior unintentionally, whether one's own behavior is or could be biased, and whether the unintentionally biased behavior has adverse consequences. Knowledge is more central to the self-concept than are automatic associations, but less central than values. Knowledge may therefore be an effective target for intervention – it is flexible enough to be influenced by new information and central enough to the self-concept to support the continuation of intervention-initiated changes.

Changing processes that are central to the self may be a necessary but not sufficient condition for producing enduring change. For truly long-term change, it may be necessary for a person to establish patterns of behavior that support the newly changed psychological processes (Miller, Dannals & Zlatev, 2017; Yeager & Walton, 2011). In this sense, the habit-breaking intervention's effects on people's tendencies to notice and label bias may have been critical for the intervention's persistence because these behaviors may have created a feedback loop with people's increased concern about discrimination – an increased tendency to

[7] Rokeach's theory is not fully consistent with modern, connectionist views of cognition (see Cox & Devine, 2015; McClelland et al., 2010). A modern connectionist view of cognition might replace the Rokeach's construct of centrality to the self-concept with the construct of the density of the connections between different aspects of the self-structure network. Regardless of which construct one uses, the implications for change processes are similar, so we maintain the "centrality" terminology for ease of communication.

notice and label the biases of others should lead to increased concern, and increased concern should also lead to noticing and labeling the biases of others.

An important implication of this analysis is that the processes that lead to the persistence of an intervention's effects are unlikely to be captured by conventional, widely-used implicit measures. Unintentional bias is a broad construct caused by a range of affective, motivational, cognitive, and behavioral processes, and responses on implicit measures only tap a small range of these processes. According to our analysis, these processes are also the ones least likely to support long-term intervention-initiated changes because they are not central to the self-concept. It may be time for intervention researchers to look beyond implicit measures by crafting interventions that target more central psychological processes and by evaluating the effectiveness of these interventions with longitudinal assessments of behaviors that contribute to recursive feedback loops.

5. Conclusion

Evidence-based interventions are needed to overcome psychological biases. The prejudice habit-breaking intervention remains a highly promising candidate for empowering people to reduce their own biases through awareness, concern, and effort. Although we did not replicate the habit-breaking intervention's effects on IAT scores, we did partially replicate its effects on discrepancies and fully replicate its effects on concern about discrimination. Moreover, change in concern seems to be both persistent and associated with change in a broad range of psychological processes related to one's orientation toward oneself and the social environment. Intervention participants were more likely to interact with Black strangers, were more likely to report noticing bias and to label it as wrong, and, two years later, were more likely to confront bias in others. Taken together, we believe this study represents promising evidence for the habit-breaking intervention's effectiveness in producing lasting psychological change.

Open practices

This article earned Open Materials and Open Data badges for transparent practices. Materials and data are available at https://osf.io/a3c8h/.

References

Apfelbaum, E. P., Sommers, S. R., & Norton, M. I. (2008). Seeing race and seeming racist? Evaluating strategic colorblindness in social interaction. *Journal of Personality and Social Psychology*, *95*, 918–932.

Banse, R., Seise, J., & Zerbes, N. (2001). Implicit attitudes towards homosexuality: Reliability, validity, and controllability of the IAT. *Experimental Psychology*, *48*, 145–160.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using **lme4**. *Journal of Statistical Software*, *67*.

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

Blair, I. V., Ma, J. E., & Lenton, A. P. (2001). Imagining stereotypes away: The moderation of implicit stereotypes through mental imagery. *Journal of Personality and Social Psychology*, *81*, 828–841.

Brewer, M. (1988). A dual-process model of impression formation. In T. Srull, & R. Wyer (Eds.), *Advances in social cognition* (pp. 1–36). Hillsdale N.J.: L. Erlbaum Associates.

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*, 365–376.

Canning, E. A., & Harackiewicz, J. M. (2015). Teach it, don't preach it: The differential effects of directly-communicated and self-generated utility–value information. *Motivation Science*, *1*, 47–71.

Carnes, M., Devine, P. G., Baier Manwell, L., Byars-Winston, A., Fine, E., Ford, C. E., ... Sheridan, J. (2015). The effect of an intervention to break the gender bias habit for faculty at one institution: A cluster randomized, controlled trial. *Academic Medicine*, *90*, 221–230.

Cialdini, R. B., Petty, R. E., & Cacioppo, J. T. (1981). Attitude and attitude change. *Annual Review of Psychology*, *32*(1), 357–404.

Cox, W. T., & Devine, P. G. (2015). Stereotypes possess heterogeneous directionality: A theoretical and empirical exploration of stereotype structure and content. *PloS One*, *10*(3), e0122292.

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

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

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

Enders, C. K., & Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: A new look at an old issue. *Psychological Methods*, *12*, 121–138.

Fiske, S. (1998). Stereotyping, prejudice, and discrimination. In S. Fiske, D. Gilbert, & L. Gardner (Eds.), (4th ed.). *The handbook of social psychology*(pp. 357–411). Boston [etc.]: The McGraw-Hill.

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

Forscher, P. S.*, Lai, C. K.*, Axt, J., Ebersole, C. R., Herman, M., & Nosek, B. A. (2017, Feb. 17). A meta-analysis of change in implicit bias. Retrieved from http://osf.io/preprints/psyarxiv/dv8tu.

Galinsky, A. D., & Moskowitz, G. B. (2000). Perspective-taking: Decreasing stereotype expression, stereotype accessibility, and in-group favoritism. *Journal of Personality and Social Psychology*, *78*, 708–724.

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.

Hulleman, C. S., & Harackiewicz, J. M. (2009). Promoting interest and performance in high school science classes. *Science*, *326*, 1410–1412.

Ibrahim, J. G., & Molenberghs, G. (2009). Missing data methods in longitudinal studies: A review. *TEST*, *18*, 1–43.

Janis, I. L., & King, B. T. (1954). The influence of role playing on opinion change. *The Journal of Abnormal and Social Psychology*, *49*, 211–218.

Kim, D. (2003). Voluntary controllability of the Implicit Association Test (IAT). *Social Psychology Quarterly*, *66*, 83.

Lai, C. K., Marini, M., Lehr, S. A., Cerruti, C., Shin, J.-E., Joy-Gaba, J. A., ... Nosek, B. A. (2014). Reducing implicit racial preferences: I. A comparative investigation of 17 interventions. *Journal of Experimental Psychology: General*, *143*, 1765–1785.

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

McClelland, J. L., Botvinick, M. M., Noelle, D. C., Plaut, D. C., Rogers, T. T., Seidenberg, M. S., & Smith, L. B. (2010). Letting structure emerge: Connectionist and dynamical systems approaches to cognition. *Trends in Cognitive Sciences*, *14*, 348–356.

Miller, D. T., Dannals, J. E., & Zlatev, J. J. (2017). Behavioral Processes in Long-Lag Intervention Studies. *Perspectives on Psychological Science*, *12*(3), 454–467. http://dx.doi.org/10.1177/1745691616681645.

Monteith, M. J. (1993). Self-regulation of prejudiced responses: Implications for progress in prejudice-reduction efforts. *Journal of Personality and Social Psychology*, *65*, 469–485.

Monteith, M. J., Ashburn-Nardo, L., Voils, C. I., & Czopp, A. M. (2002). Putting the brakes on prejudice: On the development and operation of cues for control. *Journal of Personality and Social Psychology*, *83*, 1029–1050.

Monteith, M. J., & Voils, C. I. (1998). Proneness to prejudiced responses: Toward understanding the authenticity of self-reported discrepancies. *Journal of Personality and Social Psychology*, *75*, 901–916.

Monteith, M. J., Voils, C. I., & Ashburn-Nardo, L. (2001). Taking a look underground: Detecting, interpreting, and reacting to implicit racial biases. *Social Cognition*, *19*, 395–417.

Norton, M. I., Sommers, S. R., Apfelbaum, E. P., Pura, N., & Ariely, D. (2006). Color blindness and interracial interaction: Playing the political correctness game. *Psychological Science*, *17*, 949–953.

Pettigrew, T. F. (1998). Intergroup contact theory. *Annual Review of Psychology*, *49*, 65–85.

Pettigrew, T. F., & Tropp, L. R. (2006). A meta-analytic test of intergroup contact theory. *Journal of Personality and Social Psychology*, *90*, 751–783.

Plant, E. A., & Devine, P. G. (1998). Internal and external motivation to respond without prejudice. *Journal of Personality and Social Psychology*, *75*, 811–832.

Plant, E. A., & Devine, P. G. (2003). The antecedents and implications of interracial anxiety. *Personality and Social Psychology Bulletin*, *29*, 790–801.

Plant, E. A., & Devine, P. G. (2009). The active control of prejudice: Unpacking the intentions guiding control efforts. *Journal of Personality and Social Psychology*, *96*, 640–652.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage Publications.

Richard, F. D., Bond, C. F., & Stokes-Zoota, J. J. (2003). One hundred years of social psychology quantitatively described. *Review of General Psychology*, *7*, 331–363.

Rokeach, M. (1973). *The nature of human values*. New York: The Free Press.

Smedley, B., Stith, A., & Nelson, A. (Eds.), (2003). *Unequal treatment: Confronting racial and ethnic disparities in health care*. Washington D.C.: National Academy Press.

Steffens, M. C. (2004). Is the implicit association test immune to faking? *Experimental Psychology*, *51*, 165–179.

Stephan, W. G., & Stephan, C. W. (1985). Intergroup anxiety. *Journal of Social Issues*,

41(3), 157–175.
Van Breukelen, G. J. P. (2006). ANCOVA versus change from baseline had more power in randomized studies and more bias in nonrandomized studies. *Journal of Clinical Epidemiology*, *59*, 920–925.
Wickham, H. (2009). *ggplot2: Elegant graphics for data analysis*. New York, NY: Springer.
Williams, D. R., Neighbors, H. W., & Jackson, J. S. (2003). Racial/ethnic discrimination and health: Findings from community studies. *American Journal of Public Health*, *93*, 200–208.
Yeager, D. S., & Walton, G. M. (2011). Social-psychological interventions in education: They're not magic. *Review of Educational Research*, *81*, 267–301.