Empowering People to Reduce Bias and Create Inclusion:
A Large-Scale Field Experiment Testing an Updated Bias Habit-Breaking Training

William T. L. Cox*[1,2], Emily L. Dix[1], Katharine E. Scott[1],
Xizhou Xie[1], Kristina A. Kellett[1], & Patricia G. Devine[1]

[1]Department of Psychology, University of Wisconsin – Madison
[2]Inequity Agents of Change, Inc

Author note

*Corresponding author: William T. L. Cox, Department of Psychology, University of Wisconsin – Madison, and Inequity Agents of Change Foundation.
Address correspondence to William Cox, Inequity Agents of Change Inc, 706 Vernon Ave, Madison, WI, 53714. Email: william.cox@biashabit.com.

This work was supported by two internal grants from the Office of the Provost of the University of Wisconsin-Madison awarded to the first and sixth authors, and the preparation of this manuscript was supported by Grant 1R35GM128888 from NIGMS at the NIH, awarded to the first author.

We thank Andrea Collins, Ethan Brandt, Allison Baldwin, Martha Morganstein, Saja Hakmeh, Jack Bernauer, Imani Wilson, Olivia Prager, Katie Platt, Kayla Jischke, Ash Lyke, Madison Thornton, Kendra Lange, Sage Staples, and Chloe Wigul for their work running the studies presented in this paper.

Deidentified data and materials for this project will all be publicly available, and can be found at [will be populated before publication].

This document is a preprint, and as such, the content is subject to change before publication. For the most up-to-date version of this paper, visit www.biashabit.com/research

Cite as:
Cox, W. T. L., Dix, E. L., Scott, K. E., Xie, X., Kellett, K. A., & Devine, P. G. (2022, preprint). Empowering people to reduce bias and create inclusion: A large-scale field experiment testing an updated bias habit-breaking training. *Preprint*.

Abstract

Research consistently shows that non-scientific bias, equity, and diversity trainings do not work, and often make bias and diversity problems worse (al-Gharbi, 2020; Cox & Devine, 2019; Devine & Ash, 2021; Paluck et al., 2021). One exception to this is the evidence-based bias habit-breaking training, which has successfully produced long-term reductions in bias in several randomized-controlled trials (Carnes et al., 2015; Devine et al., 2012; 2017; Forscher et al., 2017). Whereas prior, "Generation 1" versions of the habit-breaking training focused solely on a single target group or context (i.e., anti-Black bias in Devine et al., 2012; Forscher et al., 2017; anti-woman bias in STEM in Carnes et al., 2015; Devine et al., 2017), the core principles of the training involve basic cognitive processes that apply to any bias arising from stereotypes (Cox & Devine 2019; Cox et al., 2012). In the present work, we develop and experimentally test a "Generation 2" version of the habit-breaking training that empowers people to reduce biases and promote inclusion related to any stereotyped group. In Phase 1, participants were randomly assigned to complete the updated bias habit-breaking training or a control training related to environmentalism. Replicating Devine et al. (2012) with a much larger sample size (N=957), bias training participants, but not control participants, decreased in IAT bias over time, up to 6 weeks post-manipulation. Bias training participants also increased in their concern about discrimination as a serious problem, whereas controls remained the same. Phase 2 followed up with an ostensibly unrelated study 1-2 years later, in which bias training participants spoke up more than controls about bias/inclusion related topics in a mock classroom discussion. These bias/inclusion topics were not explicitly mentioned in the bias habit-breaking training program, demonstrating that bias training participants generalized and applied what they learned from the training to new stereotyped groups and inclusion-related issues. Finally, in Phase 3, participants completed a stereotyping task in another ostensibly unrelated study 2-3 years post-manipulation. Bias training participants were more likely than controls to stereotype at a lower rate, and this was significantly mediated by the previously-observed decrease in IAT bias. The present work extends prior evidence and theorizing related to the bias habit-breaking training, demonstrating considerable promise for empowerment-based approaches to diversity and bias training that engage participants as autonomous agents of change to reduce bias and create inclusion.

Empowering People to Reduce Bias and Create Inclusion: A Large-Scale Field Experiment Testing an Updated Bias Habit-Breaking Training

Increasingly, universities and other organizations have become concerned with identifying strategies to reduce rates of bias incidents, to create a more inclusive climate and to promote equity. Abundant evidence indicates that, however well-intentioned, most diversity and bias intervention efforts do not use approaches based on scientific evidence, and at best do not work and very often make bias problems worse (for excellent reviews, see al-Gharbi, 2020; Devine & Ash, 2021; Paluck et al., 2021; Pendry et al., 2007). In response, nearly every major scientific organization (e.g., NIH, NSF, AAAS) has emphasized the need for evidence-based approaches to addressing bias and promoting diversity (e.g., Moss-Racusin, et al., 2014).

Indeed, the goal of understanding, predicting, and changing human behavior is best served by the scientific method, and addressing issues of bias, diversity, and inclusion is no exception; it requires a scientific, evidence-based approach to create change and demonstrate the effectiveness of efforts to reduce bias and enhance diversity (Cox, *under review*; Cox & Devine, 2019; Devine & Ash, 2021; & Paluck, 2006). Interventions to reduce bias and improve climate related to intergroup relations should be based on a solid, evidence-based model of change. One such intervention is the *prejudice habit-breaking intervention*, (Cox & Devine, 2019; Devine et al., 2012; Forscher et al., 2017). The prejudice habit-breaking intervention is a multifaceted educational presentation, built on the prejudice habit model (Devine, 1989), which has been empirically assessed and supported in the research literature over the past 30 years (e.g., Amodio et al., 2007; Devine et al., 1991; Devine & Monteith, 1993; Monteith, 1993; Monteith et al., 2002; Plant & Devine 1998, 2009). The habit model conceptualizes prejudice and stereotyping as

"habits of mind" that can be overcome with sustained effort over time. The intervention teaches people to tune in to the potential for expressing bias unintentionally and teaches them a set of self-regulatory tools to help them overcome unintentional biases over time. Importantly, the habit-breaking training is an empowerment-based approach; it sets up the change process as being driven internally by the training recipients (Cox & Devine, 2019). The prejudice habit-breaking intervention was the first, and thus far the only intervention experimentally shown to produce long-term, lasting reductions in bias and improvements in inclusivity, with effects observed up to at least 2 years post-manipulation (Carnes et al., 2015; Devine et al., 2012; 2017; Forscher et al., 2017; for review, see Cox & Devine, 2019).

Several experimental studies have demonstrated lasting and impactful effects of the prejudice habit-breaking intervention. Participants who receive the training have shown decreases in measured implicit bias and increases in awareness of their personal vulnerability to bias and concern about bias as a serious problem (Devine et al., 2012; Forscher et al., 2017). In a follow-up study 2 years post-manipulation, intervention participants were 65% more likely than controls to speak up against bias on what they believed was a public forum than control participants.

The original version of the prejudice habit-breaking intervention focused solely on bias against Black people, but a variant of the prejudice habit-breaking intervention focused on bias against women in science, technology, engineering, and math (STEM) academic contexts (Carnes et al., 2015; Devine et al., 2017). STEM departments at UW-Madison were randomly assigned to serve as controls or receive the modified habit-breaking intervention. The habit-

breaking model of change was likewise effective in this context, leading to increased self-efficacy and awareness of gender bias, and leading to more positive departmental climates in intervention versus control departments (Carnes et al., 2015). Further, the habit-breaking training led to a 15 percentage-point increase in hiring of women faculty in intervention departments (Devine et al., 2017; Forscher, 2017). In sum, the habit-breaking intervention has experimentally shown to be effective at reducing bias and improving climate across multiple contexts (Cox & Devine, 2022).

The Present Work

An Updated Bias Habit-Breaking Training

One key limitation of the "Generation 1" habit-breaking trainings reviewed above is that each past training focused solely on bias toward one target group (i.e., toward Black people, or women in STEM). Fully addressing bias and diversity issues requires a comprehensive treatment of biases toward many different groups (e.g., Black people, Hispanic people, Muslim people, women, LGBT people). It is not feasible for universities and other organizations to implement separate interventions for every stereotyped group that may be affected by bias. Although different stereotyped groups have some unique concerns (e.g., adoption rights for same-sex couples, issues surrounding Hijab for Muslim women), core processes of how stereotypes and biases operate and lead to disparities are fundamentally similar related to different target groups (Allport, 1954; Bodenhausen & Maccrae, 1998; Cox & Devine, 2015; Cox et al., 2012; Devine, 1989; Fiske, 1998).

Our goal in the present work was to develop and test an updated version of the training that addresses intergroup biases *in general*, rather than focusing solely on one target group. This "Generation 2" habit-breaking training was adapted from the Generation 1 bias habit-breaking training focused on anti-Black bias, that was used in past work by Devine and colleagues (2012), Forscher (2016), and Forscher and colleagues (2017). The fundamental design of the Generation 2 training remained consistent with the Generation 1 trainings (Carnes et al., 2015; Devine et al., 2012; Forscher et al., 2017). Specifically, the training 1) explains the origins of stereotypes and biases and frames them as "habits of mind.", 2) teaches participants to think about reducing bias as being like breaking a habit (Devine, 1989), describing how bias habits can be broken through Motivation, Awareness, Strategies, and Effort over time, 3) explains the consequences of unintentional bias, and 4) teaches a set of evidence-based strategies to reduce bias.

The most substantial change was to broaden the potential targets of bias included in the training. This Generation 2 bias habit-breaking training was designed to encourage participants to think about the training content's applicability to any stereotyped group. The training still covered anti-Black unintentional bias, but it also included examples of stereotyping and bias based on other group statuses throughout, including other racial/ethnic groups, gender, sexual orientation, religion, nationality, and other group statuses. At several points, the training explicitly stated that the principles covered are applicable to any group about which participants have stereotypes and encourages participants to think about how to apply the concepts in the training to other target groups.

The script was further updated to reflect an array of practical and theoretical concerns, and drawing on insights from the adult learning literature, other basic research on stereotyping and bias, and models of cognitive and behavioral change (e.g., Birtel & Crisp, 2015; Cox & Devine, 2019; Cox et al., 2012; Prochaska & Velicer, 1997). See the full list of changes in Appendix A.

We evaluated the effectiveness of the training in a three-phase design (Table 1). In Phase 1, we randomly assigned participants to intervention and control conditions and measured a set of outcomes twice after the intervention, offering a chance to replicate the original test of the training (Devine et al., 2012). In Phases 2 and 3, we invited these same participants to complete ostensibly unrelated studies 1.5-2.5 years later. The Phase 2 study evaluated whether students who completed the training were more likely to speak out about bias/inclusion in a fabricated classroom discussion setting. The forms of bias/inclusion in this discussion context were not explicitly mentioned in the updated habit-breaking training, enabling us to test whether training participants generalized what they learned to new target groups, as intended. The Phase 3 study examined whether training participants were less likely to make stereotypic inferences.

Table 1. Overall Study Flow

Phase 1	Training	**In-Person Session** Random Assignment to Bias or Environmental Habit-Breaking Training
	2-3 Weeks Post-Training	**Follow-Up 1**
	4-6 Weeks Post-Training	**Follow-Up 2**
Phase 2	1-2 Years Post-Training	**Classroom Discussion Study**
Phase 3	2-3 Years Post-Training	**Stereotype Regulation Task**

Phase 1

Method

Design. First-year undergraduate students were recruited for a laboratory study and randomly assigned to one of two training conditions. Half the participants completed the Generation 2 bias habit-breaking intervention (Bias Training Condition), and the other half completed a control training. The control training closely followed the design of the bias training, but was modified to discuss breaking negative environmental habits (Environmental Training Condition). Participants who completed the in-person training session were then asked to complete a set of follow-up measures online 2 weeks after training (Follow-Up 1), and 2 weeks after they completed Follow-Up 1, they were asked to complete another set of follow-up measures (Follow-Up 2). This in-person training session and two follow-ups comprise Phase 1 of the present study.

Participants, Recruitment, and Retention. First-year undergraduate students at the University of Wisconsin-Madison were recruited via email. Every first-year student from the 2016-2017 (N = 6430) and 2017-2018 (N = 6370) academic school years who was over 18 years old was eligible and was invited at least once to participate. Participants who completed the in-person training session and both online follow-ups were compensated $40.

It was our goal to recruit as many participants as possible across the two semesters of data collection (Spring and Fall 2017). In total, 1369 participants consented, but 67 were excluded before any analyses were conducted, for abnormalities that interfered with them receiving the training content (e.g., computer errors or participants sleeping through the training).

Of the remaining 1302 participants, 345 participants did not complete any follow-up data collection sessions, leaving 957 participants (486 Bias Condition; 471 Control Condition; no retention difference by condition, chi-square = 0.0496, p = 0.82) who completed at least one follow-up (72% retention). Of these 957 participants, 53 did not complete the second follow-up session, leaving 904 participants who completed all three sessions of Phase 1.

Of the 957 participants who completed at least one follow-up session, 99% identified with the gender they were assigned at birth (66% cisgender female; 34% cisgender male), and <1% reported being transgender/intersex/genderfluid/nonbinary/genderqueer. The participants' reported races/ethnicities were 71% White, 19% Asian, 6% mixed, 2% Hispanic, 1% Black, < 1% Middle Eastern.

Procedure. As noted above, Phase 1 involved an in-person training session and two online follow-up sessions. At the in-person session, participants consented, completed baseline measures, and were randomly assigned to complete either the bias habit-breaking training (Bias Training Condition) or the environmental habit-breaking training (Environmental Training Condition). Two weeks after the in-person session, participants were emailed with a link to complete Follow-Up 1 online. Two weeks after completing Follow-Up 1, they were emailed with a link to complete Follow-Up 2. If they did not complete a follow-up within seven days of being emailed, they received a reminder email once a week until they completed the follow-up or until we had sent three reminders. Participants completed Follow-Up 1 an average of 22 days after completing the in-person session, and Follow-Up 2 was completed an average of 31 days after completing Follow-Up 1.

Environmental Habit-Breaking Training. In past tests of the habit-breaking training, participants in the control conditions did not complete a control training task; they only completed the study measures. In the present work, we wanted control participants to complete a control training task that was similar to the bias habit-breaking training in its length and level of complexity, but was unrelated to bias or stereotyping content. We therefore modified the Generation 2 bias training script into a habit-breaking training focused on changing environmentally unsustainable habits. Environmentally unsustainable behaviors (e.g., not recycling, using disposable water bottles), like unintentional biases, can be conceptualized as habits of mind. As such, changing environmental habits likewise will involve the components of the habit-breaking model (i.e., Motivation, Awareness, Strategies, and Effort). We kept the script of the environmental training nearly identical to the Generation 2 bias habit-breaking script, changing as little as possible. We altered words and examples to be about unsustainable habits, rather than unintentional biases, and taught participants strategies to develop more environmentally-friendly behaviors. Whereas the bias training uses participants' IAT score as a tool to reveal unintentional bias, we adapted a questionnaire measure, the Environmental Impact Test (EIT, described further below), to reveal and measure their levels of self-reported environmentally unsustainable behaviors.

Implicit Association Test. The Black–White evaluative IAT (Greenwald et al., 1998) was administered, as in past work (Devine et al., 2012; Forscher et al., 2017). The evaluative IAT is a dual categorization task in which people categorize pictures of Black and White faces and pleasant and unpleasant words. The logic underlying the IAT is that when two constructs (e.g., Black people and negative valence; White people and positive valence) are more closely

associated, it will be easier to categorize them when they are paired on the same response key (compatible trials). When those response key pairings are reversed (Black people and positive valence; White people and negative valence), the categorizations should be more difficult (incompatible trials). To the extent that participants have more negative and less positive automatic associations with Black compared to White people, reaction times on compatible trials should be faster compared to reaction times on the incompatible trials. These reaction times are used to compute D-scores (Greenwald et al., 2003), scored such that higher numbers indicate a stronger automatic association between Black people and negative valence. Participants in the bias training condition received feedback about their IAT score during the training program, after learning about the IAT. As in past work (Devine et al., 2012), they were told their score indicated a preference for White people over Black people, no preference, or a preference for Black people over White people, and whether this preference was strong, moderate, or slight.

Environmental Impact Test. As a measure of sustainability habits, we adapted an 18-item survey (Hunt, 2016). We chose this scale to serve a similar function to Environmental training participants as the IAT serves for Bias training participants — to give participants a general metric of how environmentally-friendly their behaviors were. We named this survey the Environmental Impact Test (EIT). Each item on the scale asks how often participants engage in a sustainable behavior (e.g., Do you switch off the light when you're the last to leave a room?"). Participants select one of three responses for each item: "Never", "Sometimes", or "Always". Each response is translated to a numerical score ("Never" is scored as 0, "Sometimes" as 1, and "Always" as 2). The original measure was designed for respondents from the U.K., so we adapted it slightly for U.S. undergraduate respondents (e.g., "bin" was changed to "trash can").

The numeric values of participants' responses were summed to create a single sustainability EIT score, with higher numbers indicating more environmentally-friendly behaviors. Participants in the environmental training condition were given their score and a short interpretation of it as part of the training program. Specifically, they were given their numeric score, and told that they were "strongly sustainable" (if their score was 31-36), "moderately sustainable" (scores of 17-30), "mildly sustainable" (scores of 9-16), or "weakly sustainable" (scores of 0-8).

Shoulds, Woulds, and Discrepancies. The discrepancy scale measures the extent to which people predict whether they would act with more prejudice than what they believe is appropriate (Monteith & Voils, 1998). Items are measured on a 1 (strongly disagree) to 7 (strongly agree) likert scale. The *Shoulds* subscale reflects people's personal standards, asking people how they believe they should act, feel or think in response to five interpersonal intergroup situations (e.g., "I should not feel uncomfortable in the company of Black people."). The items on this subscale are reverse-coded and averaged, such that higher *Shoulds* scores indicate standards that are more permissive of bias. The *Woulds* subscale reflects people's awareness of how they would actually behave, feel, or think in the same intergroup situations (e.g., "I would feel uncomfortable if I were the only White person in a group of Black people."). Higher *Woulds* scores indicate participants have higher levels of awareness that they would respond in biased ways. The *Discrepancies* score is computed by subtracting the *Shoulds* score from the *Woulds* score. Higher *Discrepancies* indicate that people are aware that they would behave with more bias than they believe they should.

Concern About Discrimination. The *Concern* scale comprises four items evaluating participants' beliefs that racial discrimination is a serious social problem (Devine et al., 2012). Each item (e.g., "I consider racial discrimination to be a serious social problem") is scored using a 1 (strongly disagree) to 10 (strongly agree) scale. Items are coded such that higher numbers indicate greater concern for racial bias as a serious problem and averaged into a single *Concern* score.

Evaluation Items. Ten evaluation items about participants' perceptions of the training were also collected during the two follow ups. Each item (e.g., "I think the intervention positively affected my life") was scored using a 0 (strongly disagree) to 100 (strongly agree) slider. For analyses, we averaged participants responses across the two Follow-Ups, because they did not differ over time. The full list of items appears in a table in the Results section.

Procedural Shift. It was our initial intention to collect the key measures (IAT, EIT, Shoulds, Woulds, and Concern) at three timepoints (Baseline, Follow-Up 1, and Follow-Up 2). When data collection began, participants completed that full battery of measures, then were immediately randomly assigned to receive either the bias or the environmental habit-breaking training. Some participants mentioned to, or in front of, our experimenters that they believed the training they completed was chosen based on their responses to the baseline measures. For example, after one study session, an experimenter heard a discussion between two participants who knew each other but had been randomly assigned to different conditions. After telling one another what their respective trainings were about, one commented that he was surprised he got the bias training whereas his friend got the environmental training, because they both "knew the

friend was more racist and needed bias training more." A few of these instances raised concerns that it maybe be a problem to have so many bias-related measures in the same session, directly before the training. Based on these concerns, we decided to change the study procedure, about one-third of the way through data collection. We eliminated most of the bias-related baseline measures for the remainder of data collection. After this procedural shift, at baseline, all participants completed some generic filler measures and the EIT. Only participants who were in the bias training condition completed the IAT at baseline. To match the time taken up by the IAT for bias condition participants, participants in the environmental training condition completed some additional filler surveys related to campus life but unrelated to the present study. We hoped that reducing the amount of bias-related baseline measures would reduce potential demand characteristics or other potential confounds between conditions.

We made this procedural change without looking at any of the data collected up to that point. We decided that our primary data analytic strategy would be to look at condition differences in the two follow-up sessions, without comparison to baseline.

Results and Discussion

Analytic Approach. Because of the procedural shift described above, we decided *a priori* that our primary analytic approach would be to conduct 2 (Training Condition: Bias vs. Environmental) x 2 (Timepoint: Follow-Up 1 vs. Follow-Up 2) mixed ANOVAs, with Training Condition between-subjects and Timepoint within-subjects. Because of the potential impact of changing procedure partway through data collection, we conducted additional analyses testing whether the procedural shift altered our observed patterns. Specifically, we conducted additional

2 (Condition) x 2 (Timepoint) x 2 (Procedural Shift: Before vs. After) mixed ANOVAs for each outcome. The procedural shift did not interact with any of the reported effects, all p's > 0.2.

One limitation of this planned analytic approach, however, is that it drops participants with missing data, thus can only include with the 904 participants who completed both follow-up sessions. This sample size of 904 granted us 1- β of 0.96 power for repeated measures correlated at 0.35 to detect even a small between-subjects effect size of f = 0.1 (d = 0.2). Because we committed to this data analytic approach before conducting analyses, those will be reported as the primary hypothesis tests. After primary data analysis, we conducted additional analyses using linear mixed effects modeling (LMEM), because LMEMs enabled us to retain participants with missing data. The LMEMs allowed us to 1) include participants who didn't complete the second follow-up, and 2) allow us to include baseline measures for the participants who completed them. The patterns of results did not differ between these two analytic approaches, and we report both approaches. For key analyses, we supplement traditional null hypothesis significance testing with Bayes factor analyses.

Replication Analyses. Of first interest was whether the new Generation 2 of the bias habit-breaking training replicated the key effects of the Generation 1 trainings. The three key effects of the Generation 1 training were that, over time, compared to control participants, the bias training participants showed 1) decreases in implicit bias as measured by the IAT, 2) increases in Concern, and 3) increases in Woulds and Discrepancies (Devine et al., 2012). See descriptive statistics in Table 2 and bivariate correlations in Table 3.

Table 2

Means and standard deviations of implicit and explicit variables by intervention condition for each follow up.

		Bias			Environment		
		N	Mean	SD	N	Mean	SD
Follow up 1	IAT	480	0.26	0.43	456	0.34	0.39
	Shoulds	470	1.51	0.78	467	1.53	0.80
	Woulds	470	3.30	1.28	467	3.27	1.21
	Discrepancies	470	1.79	1.28	467	1.73	1.22
	Concern	470	7.38	1.82	467	7.39	1.94
Follow up 2	IAT	448	0.29	0.44	447	0.36	0.40
	Shoulds	439	1.56	0.82	446	1.57	0.88
	Woulds	439	3.19	1.25	446	3.22	1.25
	Discrepancies	439	1.63	1.21	446	1.64	1.29
	Concern	439	7.56	1.88	447	7.36	2.01

Table 3

Correlations between implicit and explicit variables within intervention conditions. Correlations in the bias condition are shown above the diagonal; correlations in the environmental condition are shown below the diagonal.

	IAT[B]	IAT[FU1]	IAT[FU2]	Shoulds[FU1]	Shoulds[FU2]	Woulds[FU1]	Woulds[FU2]	Discrepancies[FU1]	Discrepancies[FU2]	Concern[FU1]	Concern[FU2]
IAT[B]	–	0.28	0.21	0.09	0.13	0.23	0.22	0.17	0.14	-0.16	-0.13
IAT[FU1]	0.31	–	0.37	0.04	0.09	0.13	0.12	0.1	0.06	-0.13	-0.07
IAT[FU2]	0.28	0.31	–	0.04	0.02	0.08	0.03	0.06	0.01	-0.09	-0.06
Shoulds[FU1]	0.13	0.08	0.06	–	0.52	0.31	0.33	-0.3	-0.01	-0.33	-0.39
Shoulds[FU2]	0.1	0.03	0.03	0.58	–	0.32	0.37	0	-0.3	-0.33	-0.39
Woulds[FU1]	0.24	0.2	0.25	0.31	0.27	–	0.78	0.82	0.59	-0.35	-0.38
Woulds[FU2]	0.19	0.18	0.16	0.27	0.3	0.73	–	0.59	0.78	-0.35	-0.38
Discrepancies[FU1]	0.14	0.15	0.2	-0.34	-0.11	0.79	0.54	–	0.6	-0.15	-0.14
Discrepancies[FU2]	0.11	0.15	0.13	-0.13	-0.39	0.52	0.76	0.6	–	-0.14	-0.13
Concern[FU1]	-0.19	-0.16	-0.12	-0.33	-0.41	-0.38	-0.34	-0.16	-0.05	–	0.77
Concern[FU2]	-0.26	-0.17	-0.14	-0.35	-0.4	-0.36	-0.36	-0.13	-0.08	0.82	–

FU1: Follow up 1; FU2: Follow up 2

IAT. A 2 (Condition: Bias Training vs Environmental Control Training) x 2 (Timepoint: Follow-Up 1 vs Follow-Up 2) mixed ANOVA revealed a main effect of Condition, such that participants in the bias training condition had lower scores on the IAT than the environmental control participants across both follow ups, $F(1,883) = 10.14$, $p < .001$. There was no main effect of Timepoint, $F(1,883) = 0.88$, $p = 0.348$ and no Condition x Timepoint interaction, $F(1,883) = 0.03$, $p = 0.872$.

Bayes Factor analyses conducted using JASP replicated the null hypothesis significance testing approach. A Bayesian repeated measures ANOVA (Cauchy = 0.707) indicated that the best model to account for the data is a model that includes the main effect of Intervention. $BF_M = 21.937$ suggests that the data is 21.937 times more likely to occur under Intervention main effect model than any other models, including the null. Following Lee and Wagenmaker's (2014) rule of thumb, this BF indicates that there is *strong* evidence in favor of the hypothesis. See Table 4.

Table 4. Bayes factor model tests for IAT effect

Model Comparison ▼

Models	P(M)	P(M|data)	BF_M	BF_{10}	error %
Intervention	0.200	0.846	21.937	1.000	
Null model (incl. subject)	0.200	0.077	0.332	0.091	4.966
RM Factor 1 + Intervention	0.200	0.066	0.283	0.078	5.274
RM Factor 1	0.200	0.007	0.026	0.008	5.996
RM Factor 1 + Intervention + RM Factor 1 ✻ Intervention	0.200	0.005	0.020	0.006	6.776

Note. All models include subject

Because all bias condition participants completed the IAT at baseline, we were able to conduct an additional within-condition analysis to further confirm that bias training participants *decreased* in IAT bias. We conducted a one-way repeated-measures ANOVA on the bias training

participants only, comparing their IAT scores at baseline and the two followups. This repeated measures ANOVA revealed that the bias condition participants' IAT scores did in fact decrease, $F(2,886) = 12.48$, $p < 0.001$. Tukey post-hoc tests revealed decreased implicit bias from baseline to Follow-Up 1, t (886) = 4.6, $p < 0.001$ and remained diminished at Follow-Up 2 compared to baseline, t (886) = 4.0, $p < 0.001$. A one-sided Bayesian Paired Samples T-test (Cauchy = 0.707) tested the alternate hypothesis that the bias training participants' IAT scores in the Follow-Up 2 were lower than those before the training. The estimated Bayes Factor, $BF_{-0} = 141.926$, indicates that these data provide *extremely strong* evidence in favor of the hypothesis that the bias training participants' IAT scores decreased from before training to last follow up session. Similarly, we performed the same one-sided Bayesian Paired Samples T-test for the subset of control participants for whom we had baseline IAT scores. The Bayes Factor, $BF_{-0} = 0.067$, and $1/BF_{-0} = 14.930$, indicates the data were 14.930 more likely to occur under the null hypothesis. This BF indicates there is strong evidence in favor of the null hypothesis that IAT scores were not affected for control participants.

Lastly, we conducted a LMEM using lme4 (Bates, Mächler, Bolker, & Waler, 2015), which does not drop participants case-wise, enabling us to include participants who only completed one follow-up. This LMEM also allows us to include the baseline IAT data we had for all bias training participants and some of the environmental training condition participants. The LMEM included a random intercept for each participant, random slopes for measures that varied within participants, and the correlation between the intercepts and slopes. The model included linear and quadratic effects of time, as well as indicators for condition and the condition by time interactions. The model also contained random slopes for time.

This LMEM revealed a main effect of Condition, such that, on average across time points, the bias training participants showed reduced implicit bias compared to environmental control participants, $F(1, 1054.3) = 4.83$, $p = 0.03$. This main effect of Condition interacted with linear Time, $F(1, 1813.2) = 7.36$, $p = 0.007$, such that participants did not differ in IAT at baseline, but bias training participants' IAT scores decreased in the follow-ups, while control participants' IAT scores remained the same. See Figure 1.

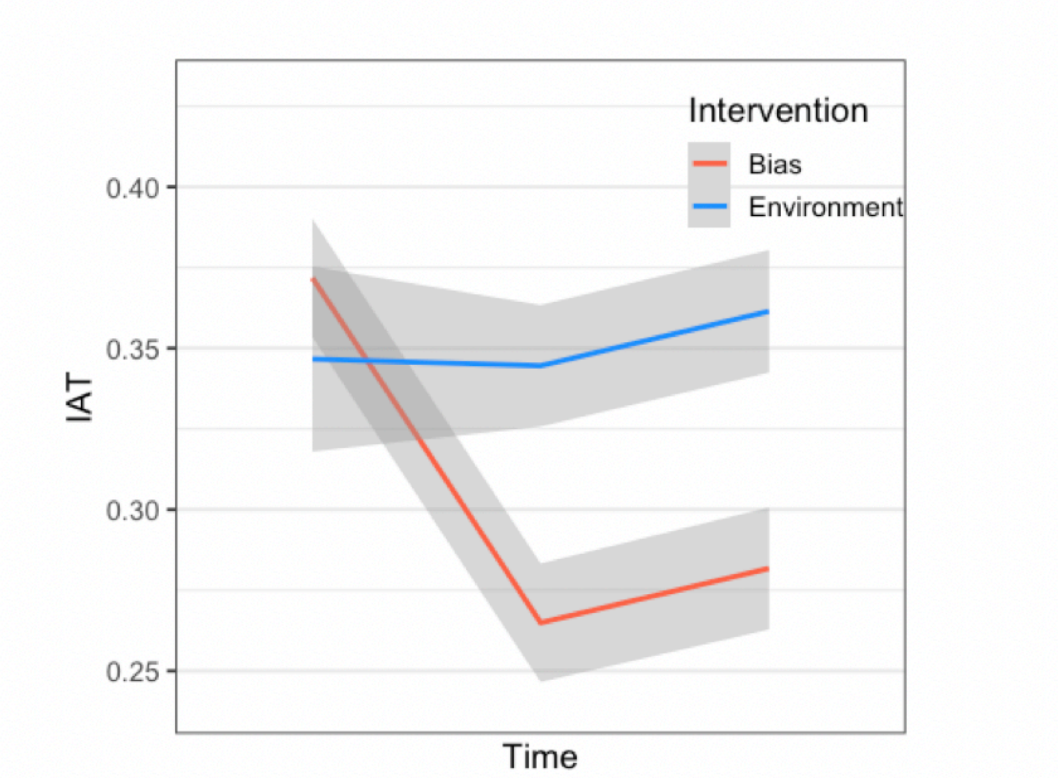

Figure 1. LMEM Point Estimates for IAT Effects with SE Envelope

Concern, Should, Woulds, and Discrepancies. The 2 (Condition) x 2 (Timepoint) mixed ANOVA for Concern revealed a Condition x Timepoint interaction, F (1, 884) = 6.67, p = 0.01, showing that bias training participants increased in Concern over time compared to control participants, with higher levels of Concern at Follow-up 2. This pattern partially replicates patterns of past work with regard to the Concern measure (Devine et al., 2012; Forscher et al., 2017), with bias training participants increasing in Concern relative to control participants. This Concern pattern does not *fully* replicate the prior work, however, because the original Generation 1 study (Devine et al., 2012) saw increases in Concern at the first follow-up, whereas in the present work, the increase only appears at the second follow-up.

There were no effects of Condition on Shoulds, Woulds, or Discrepancies, all p's > 0.3. These patterns also fail to replicate prior work with Generation 1 of the habit-breaking training, which saw increases in Woulds and Discrepancies. The seeming non-replication patterns of Concern, Woulds, and Discrepancies make more sense, however, when comparing the means from the present study and the original Devine and colleagues (2012) study. In the original study, bias training participants' mean level of Concern was 6.08 at baseline, and increased to 6.42 at Follow-Up 1 and 6.57 at Follow-Up 2. In the present study, for the subset of participants for whom we have baseline Concern, their levels of Concern at baseline are 7.36 (Bias training condition) and 7.41 (Environmental training condition) — higher than the highest levels of Concern observed in the original study.

A similar pattern occurs for Woulds and Discrepancies. In the original study, bias training participants' mean level of Woulds (and Discrepancies) were 3.16 (1.17) at baseline, and

increased to 3.51 (1.63) at Follow-Up 1 and 3.38 (1.49) at Follow-Up 2. For the present study, the participants who completed baseline measures have Woulds (Discrepancies) that are higher than the highest levels observed in the original study, for both bias training condition participants 3.67 (2.30) and environmental training condition participants 3.71 (2.35). The Woulds and Discrepancies did not increase as a function of the Generation 2 training, but the participant population's scores were already higher than the final Would/Discrepancy scores the Generation 1 training was able to yield, perhaps suggesting a potential ceiling effect on these measures.

Evaluation. Participants in both conditions completed evaluation items, but of key interest was bias participants' evaluations of this new Generation of the bias habit-breaking training. One of our goals for the revised intervention was to make sure the content was relevant and applicable to participants who are members of stigmatized groups, most especially nonWhite people.

See Table 5 for descriptive statistics of nonWhite and White participants' responses to each evaluation item, and one-sample t-tests and Bayes factors comparing the average responses to the scale midpoint, 50, which gives us an indicator of whether they were significantly more favorable than the neutral midpoint. Overall, nonWhite participants had very favorable responses to the training. They were happy they went through the training both immediately after (M=70.16, sd = 23.10) and several weeks later (M = 72.73, sd = 23.17). They learned things they did not know (M= 61.95, sd = 29.30), and think it would be good for all incoming students to take this intervention (M = 67.87, sd = 25.25).

Table 5

Evaluation Items, Descriptives, and Comparisons to Scale Midpoint

	nonWhite		White		All Bias Training Participants		
Items	Mean	Sd	Mean	Sd	One-Sample t >50	*p*	BF (> 50)
Immediately after completing the intervention, I was happy I went through it.	71.00	21.57	66.07	23.26	16.27	<0.001	7.80E+43
Now, several weeks after completing the intervention, I am happy I went through it.	73.20	21.83	69.93	22.69	19.70	<0.001	2.05E+59
The intervention was too long.	58.97	23.91	54.27	25.42	4.89	<0.001	11440.78
The intervention's content was valuable.	70.63	22.48	68.29	22.95	17.76	<0.001	5.09E+50
I learned things I did not know.	60.79	27.08	55.26	26.96	5.50	<0.001	224899.75
I think the intervention positively affected my life.	65.54	21.45	62.99	22.77	13.13	<0.001	3.56E+30
I think it would be good for all incoming students to take this intervention.	68.18	22.49	58.25	27.13	9.27	<0.001	7.79E+15
I've found the intervention content helpful in my daily life.	59.43	23.05	51.16	26.22	3.15	<0.001	13.67
I would recommend this intervention to a friend.	61.17	24.36	53.23	27.59	4.55	<0.001	2439.62
I have discussed things I learned in the intervention with other people.	53.95	28.26	47.54	30.66	-0.31	0.62	0.042

Environmental Impact Test. Although the Environmental Habit-Breaking Training was developed primarily as a control task, its fundamental model of change should, in theory, be effective at equipping participants to change their habits related to sustainability and the environment. To test whether the environmental habit training was effective at helping participants change their self-reported environmental behaviors, we conducted a 2 (Condition:

Environmental Training vs. Bias Training) x 3 (Timepoint: Baseline vs Follow-Up 1 vs Follow-Up 2) mixed ANOVA on participants' EIT scores. This analysis revealed main effects of Condition, F (1, 902) = 10.80, $p < 0.001$ and Timepoint, F (2, 1804) = 10.80, $p < 0.001$ that were qualified by their interaction, F (2, 1804) = 5.09, $p = 0.006$, $BF_M = 4.45$. The interaction revealed that participants in both conditions started at the same level of EIT, but the environmental training participants increased over time, indicating that they reported engaging in more environmentally-friendly behaviors over time. See Figure 2.

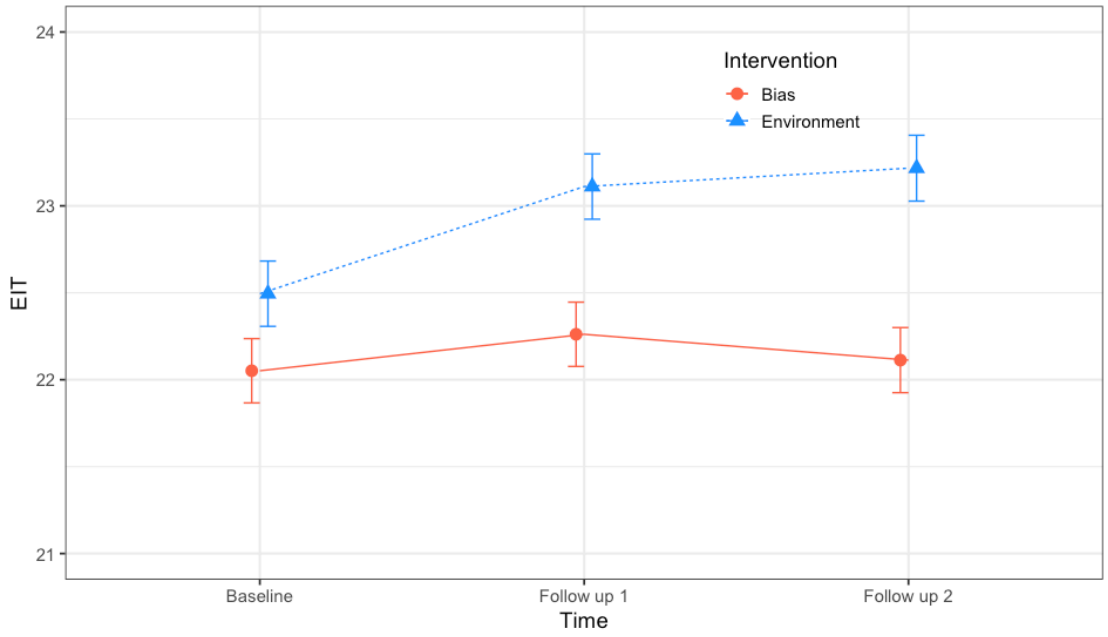

Figure 2. EIT Scores over time

Phase 1 General Discussion. Replicating prior work with Generation 1 versions of the habit-breaking training, our updated Generation 2 version of the training caused bias training participants, but not control participants, to have lasting decreases in IAT bias. Bias training participants also increased in Concern. Although the Environmental training was developed primarily as a control task, it was encouraging to see that it led to increases on the EIT, a measure of participants' practical, day-to-day sustainability behaviors. Translating the habit-breaking intervention to the sustainability domain provides some further validation of the general habit-breaking model of change.

Three key measures from past work testing Generation 1 of the bias training, Concern Woulds, and Discrepancies, had much higher baseline levels in the present study than in the original study (Devine et al., 2012). Although we have no direct evidence about the reason for these baseline differences in Concern, Woulds, and Discrepancies, we speculate that the population was more concerned and aware of unintentional bias, due to greater discussion of bias in the public arena. The data collection for the original test of Generation 1 of the habit-breaking training was conducted in 2008, and the data for the present study were collected in 2017, very soon after the 2016 election, which involved many discussions related to race and bias. The timing of the present study also coincided with many incidents related to racial biases that received national attention (e.g., Stevenson, 2017), and a few incidents on the campus where the study data were collected (The Associated Press, 2017). We believe it is a reasonable speculation that various discussions on campus and in the public arena raised participants' concern and awareness about bias, leaving less room for the Generation 2 training to create further increases. At Follow-up 2, we did see an increase in Concern for bias training participants, above the

already-heightened levels observed at baseline, making this change even more impressive with this sample.

Two key goals of the training are to empower recipients to reduce bias and to promote inclusion. Phase 1 provides evidence that the Generation 2 training replicates the Generation 1 trainings in meeting the goal of reducing bias, at least as measured by the IAT. Bias training participants show reduced bias, as measured by the IAT, and increased Concern, which is one indicator of efforts to reduce bias and promote inclusion (Dix, Harris, & Devine, *under review*). In Phase 2, we extend our test of the longevity of the Generation 2 training's effects, and assess its effects on additional bias and inclusion related behaviors.

Phase 2

The bias habit-breaking training is designed to equip participants with a better understanding of bias, how it impacts others, and how to address it to create a more inclusive social environment. In prior work, participants who completed the Generation 1 bias habit-breaking training were 65% more likely than control participants to speak up about bias in a public forum, two years after the experimental manipulation (Forscher et al., 2017). If the training is successful at empowering recipients to reduce bias and create inclusion, we would expect them to be more likely to speak up about issues related to bias and inclusion. We sought to assess these outcomes in Phase 2.

As noted earlier, the largest update in Generation 2 of the bias habit-breaking training was encouraging participants to generalize the content to various stereotyped groups. In Phase 2, we sought to test whether these updates were effective at encouraging participants to speak up

about bias related to groups other than Black people (as was assessed in Forscher et al., 2017). The training is also designed to equip participants with tools that enable them to apply the training's concepts to other forms of bias not covered by the training. Ideally, one outcome of the bias habit-breaking training would be that participants are equipped to recognize and effectively address new forms of bias or inclusion and speak up about them, as needed, even if the training gave them no formal instruction about those forms of bias or inclusion. As a test of this aspect of the training, Phase 2 provided participants with two opportunities to speak up about bias/ inclusion related to two social groups that were never mentioned in the training: Arab-Muslim people being targeted for extra screenings by the United States Transportation Security Administration (TSA), and the need for gender neutral bathrooms, especially for gender nonbinary people. At no point did the training explicitly mention anything related to either of these issues. Any differences in speaking up about bias/inclusion related to Arab-Muslim or gender nonbinary identities as a function of condition, therefore, would provide evidence that bias training participants are indeed generalizing what they learned in the training, beyond the types of bias the training explicitly discussed.

We gave control and bias training participants opportunities to speak up about bias/ inclusion in a context that was highly relevant for our college student population: classroom discussions. In a mock online classroom discussion 1-2 years post-training, participants read and discussed three news articles. One article served as a bias-unrelated control, one article related to Anti-Muslim bias, and the third article related to Gender Nonbinary bias. We used quantitative text analysis metrics to assess the extent to which bias training participants and control

participants spoke up about bias or inclusion-related topics in response to the two bias-related articles.

Method

Participants and Design. All participants from Phase 1 who completed at least one follow-up session were invited to participate via email. Before collecting any data, we computed an a priori power analysis using G*Power, to have 1-beta = 0.8 power to detect an estimated effect size of d = 0.4 for an independent samples t-test comparing two groups. This power analysis indicated a sample size of 100 participants per condition. We surmised that the study content related to Muslim people and queer or gender non-binary people might elicit different reactions from participants connected to those groups, or other stigmatized groups who might be more familiar with bias and inclusion issues in general. Because of this possibility, we set our recruitment cutoff based on the number of White, straight, non-Muslim, cisgender participants recruited. We did not exclude participants based on these demographics; we merely continued data collection until the recruitment goal was reached counting only the White, non-Muslim, cisgender participants. These demographics did not interact with any of the effects reported (p_{race} = 0.47, p_{muslim} = 0.27, p_{trans} = 0.85), so they are discussed no further. Phase 2 recruited 304 participants in total (68.09% Female, 73.35% White, 50.65% Bias condition). Participants were compensated with a $20 Amazon Gift Card in exchange for their participation.

Procedure. We recruited participants via email. The recruitment emails purportedly came from the university First-year Interest Group (FIG) committee. FIG is a program consisting of clusters of three UW-Madison courses, linked together to explore a common theme, and offered

to incoming freshmen who attend these classes together as a cohort. Participants thought they were randomly selected to a study that assessed how students engaged in online discussion forums in Canvas, the online learning management system used at their university. Participants consented and were told that the study was focused on online class discussions.

Five article titles were presented to participants, with a short description of each article. The topics covered by the articles were: gun control in the U.S., gender neutral bathrooms, cost of higher education, airport security, and internet privacy. Including three non-bias/inclusion related articles masked the purpose of the study. Participants rated their interest on each topic using a 5-point Likert scale (from 1 = “Not at all Interested” to 5 = “Extremely Interested”), and then ranked the five articles in order of their level of interest in reading them. There was no significant difference between Conditions on the interest or rankings of articles, so they are discussed no further. Participants were told that three of these five articles were randomly selected for them to read. In reality, all participants received the same 3 articles, in this order: Internet Privacy (Control), Airport Security (Muslim Bias), and Gender Neutral Bathrooms (Gender Nonbinary Inclusion).

Similar to common class assignments that involved posting a discussion comment and responding to classmates’ comments, participants were first asked to write 3-5 sentences about their thoughts on the article. They were then shown five comments ostensibly written by previous participants, and they were asked to choose two of these comments and to write a short reply to each of the two comments they chose.

Materials. Each of the three articles was a real article retrieved from a reputable news source. The first article participants saw was the Control article, which discussed internet use and privacy. This control article familiarized participants with the procedure, and provided a comparison for any patterns observed in the bias-related articles. The second article was the Muslim Bias article, and it pertained to Muslim people's experiences facing additional screening procedures by the TSA. It was entitled "Traveling while Muslim Complicates Air Travel." The third article was the Gender Nonbinary Inclusion article, and it pertained to gender-neutral bathrooms. It was entitled "Why All Bathrooms should be Gender-Neutral" and shared an opinion supporting gender-neutral bathrooms. This article, written from the perspective of gender nonconforming person, discussed inclusivity concerns related to gender nonbinary people, which was a potentially unfamiliar target group to participants.

Participants saw five comments, ostensibly written by other students, in response to each article. For the two bias/inclusion-relevant articles, these stimulus comments contained a mix of mundane content (e.g. "I enjoyed reading the article"), content denying bias/inclusion concerns (e.g. "I don't see how [TSA agents] are being "biased"), and content acknowledging the validity of bias/inclusion concerns (e.g. "I can really see how much of an impact [gender-neutral bathrooms] would have for the LGBTQ+ community to have gender-neutral bathrooms."). Together, the sets of five comments provided a range of content that participants could respond to in their two replies.

Quantitative Text Analysis and Dictionary Validation. Participants' writings were quantitatively analyzed using the Quanteda package in R. First, all responses were spell-checked

and corrected using the automated spell-check function within Google Sheets. Following the standard recommendations of the developers of the Quanteda package, the text analysis program removed the default English stopwords (e.g., “a”, “and”, “me”) and punctuation. A dictionary of 41 bias/inclusion-related words (e.g., “bias”, “racism” “fair”) was created to assess the frequency of bias/inclusion-related words used in participant responses. See Appendix B for the full list of words and further details of the quantitative text analysis.

To test the validity of the quantitative bias/inclusion dictionary, we assessed the extent to which the quantitative scores related to human-coded impressions of the participants’ written responses. All participants’ responses were independently coded by two research assistants blind to participants’ condition. The full coding scheme and details about these variables are available in Appendix B. Each response was coded by two research assistants on four key dimensions: whether the participant took a clear Anti-Bias Stance, whether they demonstrated Perspective-Taking, the conceptual Thoughtfulness of their response, and whether they mentioned core Training Concepts (e.g., “habit”, “implicit bias”). Inter-rater reliability kappa was 0.69. A third independent coder resolved any coding difference produced by the first two coders.

As a validation of the quantitative bias-related dictionary, we conducted Spearman rank correlations with the total bias/inclusion-related word counts and the means of the participants’ human-coded variables. When the quantitative text analysis program indicated that participants used more bias/inclusion-related words, the human coding indicated that those participants were more likely to take a strong Anti-Bias Stance ($r_s(304) = 0.38$, $p < 0.001$), to demonstrate Perspective Taking, ($r_s(304) = 0.37$, $p < 0.001$), to have higher Thoughtfulness in their responses,

($r_s(304) = 0.57$, $p < 0.001$), and to mention Training Concepts ($r_s(304) = 0.14$, $p = 0.013$). The quantitative bias/inclusion-related word count also correlated positively with participants' level of Concern at Follow-Up 2, from Phase 1 ($r = 0.24$, $p < 0.001$). These consistent patterns support our use of this text analysis dictionary as a quantitative metric reflecting participants' engagement with and discussion of bias, inclusions, and related issues.

For full details of the dictionary development, human coding, and to see the full dictionary, refer to Appendix B.

Results

Length of Responses. We first assessed overall length of responses, to see whether bias training participants wrote more overall on in their responses to the two bias-related articles. LMEMs comparing conditions showed that bias training participants indeed wrote more than control participants, F (1, 301.98) = 4.95, $p = 0.03$. On average, bias training participants wrote 14 more words in their three responses to the Muslim bias-related article, and 12 more words in their three responses to the Gender Nonbinary bias-related article. We interpret overall length as an indicator of engagement; participants who have been through the bias training have an ongoing investment in bias/inclusion-related issues, and thus have more to say and are more willing to take the time to say it. Of course, greater length could merely reflect verbosity — using more words without adding more, or more valuable, content (although, it is unclear why completing a bias training ~1.5 years prior to this Phase would have caused greater empty verbosity). Response length for the control article helps address this concern: For the control article, there was no significant difference in length between conditions, F (1, 302.05) = 1.69, p =

0.19, BF_{+0} = 0.50. Also, for the two bias-related articles, overall response length correlates positively with the human-coded Thoughtfulness of their responses, $r_s(304) = 0.87$ $p < 0.001$, indicating that greater length likely reflects participants making more complex points and bringing in more relevant insights.

Bias/Inclusion Dictionary Text Analyses. We first assessed whether bias training participants were more likely to speak up about bias/inclusion topics in general. For each of the three writings related to the two bias-related articles, the quantitative text analysis dichotomously identified whether each writing discussed bias/inclusion at all, or did not discuss bias/inclusion. In their original comments about the articles, there was no difference in speaking up about bias/inclusion between bias training and control participants (B = -0.20, beta = -0.25, p = 0.331, OR = 0.82). In their replies to the generated comments, however, bias training participants were 52% more likely to speak up about bias/inclusion than control participants (B = 0.42, beta = 0.44, p = 0.011, OR = 1.52).

We next tested whether participants differed in *how much* they discussed bias/inclusion-related topics, as a function of training condition. To identify the correct model for this phase of the quantitative text analysis, we subjected the bias/inclusion-related word count to a likelihood ratio test between a Poisson and a negative binomial model. Results indicated the count data should be fit using a negative binomial model ($p < 0.001$). Analysis was conducted using mixed effects models using lme4 (Bates et al., 2015). Participants' three written responses to each article (their initial comment and two replies) were treated as repeated measures in a Generalized Linear Mixed-Effects Model with Condition (Bias training vs Environmental training control) as

a between-subjects effect, with participant as a random effect. The model used dummy codes to compare participants' responses to the Muslim Bias article and the Gender Nonbinary Bias article to their responses to the Control Article. See Table 6.

Table 6. LMEM for Counts of Bias/Inclusion Words

Fixed effects	β	B	P	IRR	95% CI IRR
Intercept		-3.28	< 0.00001	0.04	(0.03,0.05)
Condition	0.35	1.03	0.0035	2.81	(1.41, 5.64)
Muslim article (vs Control)	1.15	3.60	< 0.00001	36.74	(25.93, 52.06)
Gender Nonbinary article (vs Control)	1.16	3.62	< 0.00001	37.41	(26.40, 53.01)
Condition X Muslim article (vs Control)	-0.17	-0.85	0.017	0.43	(0.21, 0.86)
Condition X Gender Nonbinary article (vs Control)	-0.18	-0.93	0.0086	0.39	(0.20, 0.79)

The main effect of Condition revealed that bias training participants talked about bias/ inclusion-related topics more than control participants (B = 1.03, beta = 0.35, p = 0.0035, IRR = 2.81). The IRR indicates that bias training participants overall talked about bias 181% more than control participants. A Bayesian independent sample t-test further supported this conclusion, Cauchy = 0.707, BF_{+0} = 3.589, indicating that these data provide moderate evidence in favor of our hypothesis.

It is important, however, to test whether participants are using more bias-related words *when bias is relevant*, rather than just using more of them all the time. The main effects of article type reveal that people discussed bias topics more when they were relevant, in their responses to the Muslim Bias article compared to the control article B = 3.60, beta = 1.15, $p < 0.0001$, IRR = 36.74, and in their responses to the Gender Nonbinary Bias article compared to the control article, B = 3.62, beta = 1.16, $p < 0.0001$, IRR = 37.41. Each of these main effects were qualified

by their interactions with Condition, such that Bias training participants discussed bias topics more than control participants for the Muslim Bias Article B = -0.85, beta = -0.17, p = 0.017, IRR = 0.43 and the Gender Nonbinary Bias Article B = -0.93, beta = -0.18, p = 0.0086, IRR = 0.39 compared to the control article. In summary, participants who had received the bias training 1-2 years ago were more likely to talk about bias and use bias-related words, when bias was relevant to the classroom discussion context. See Figures 3 and 4.

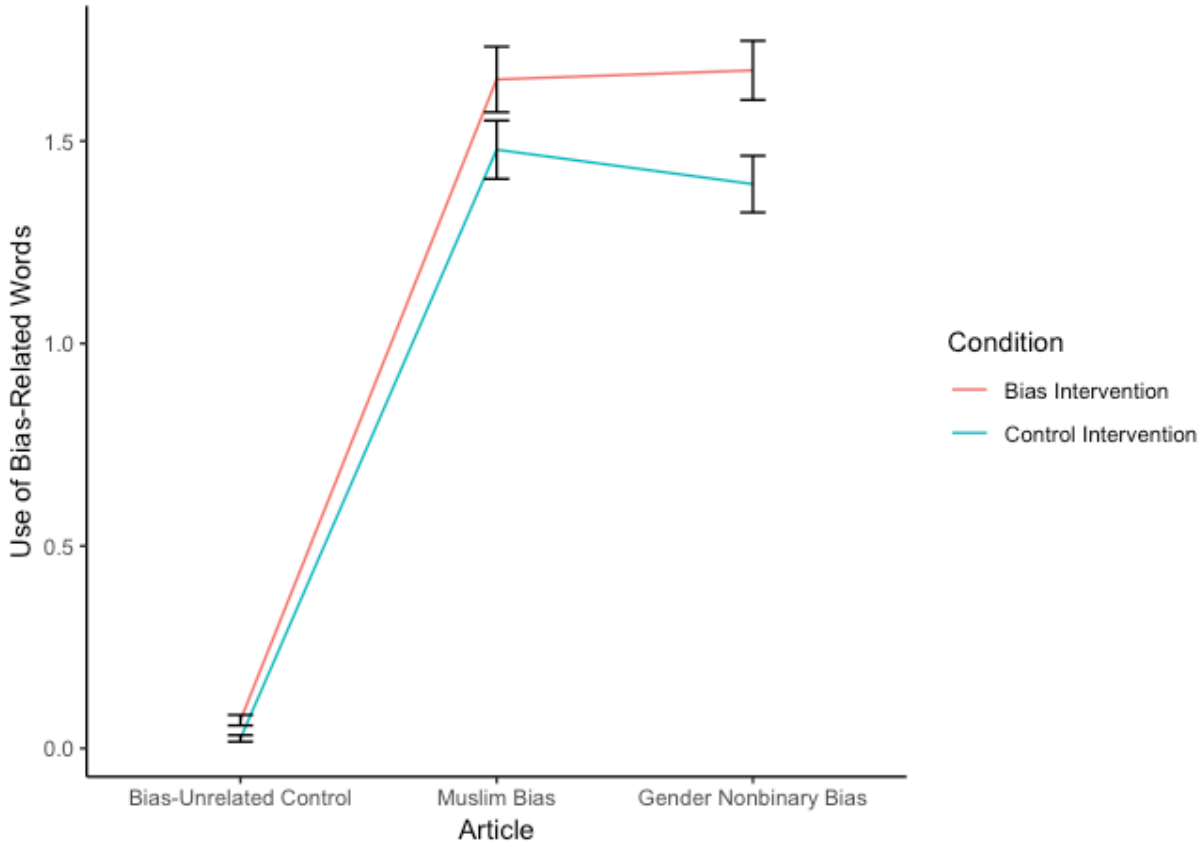

Figure 3. Mean Number of Bias/Inclusion Related Words Per Writing, By Article and Condition.

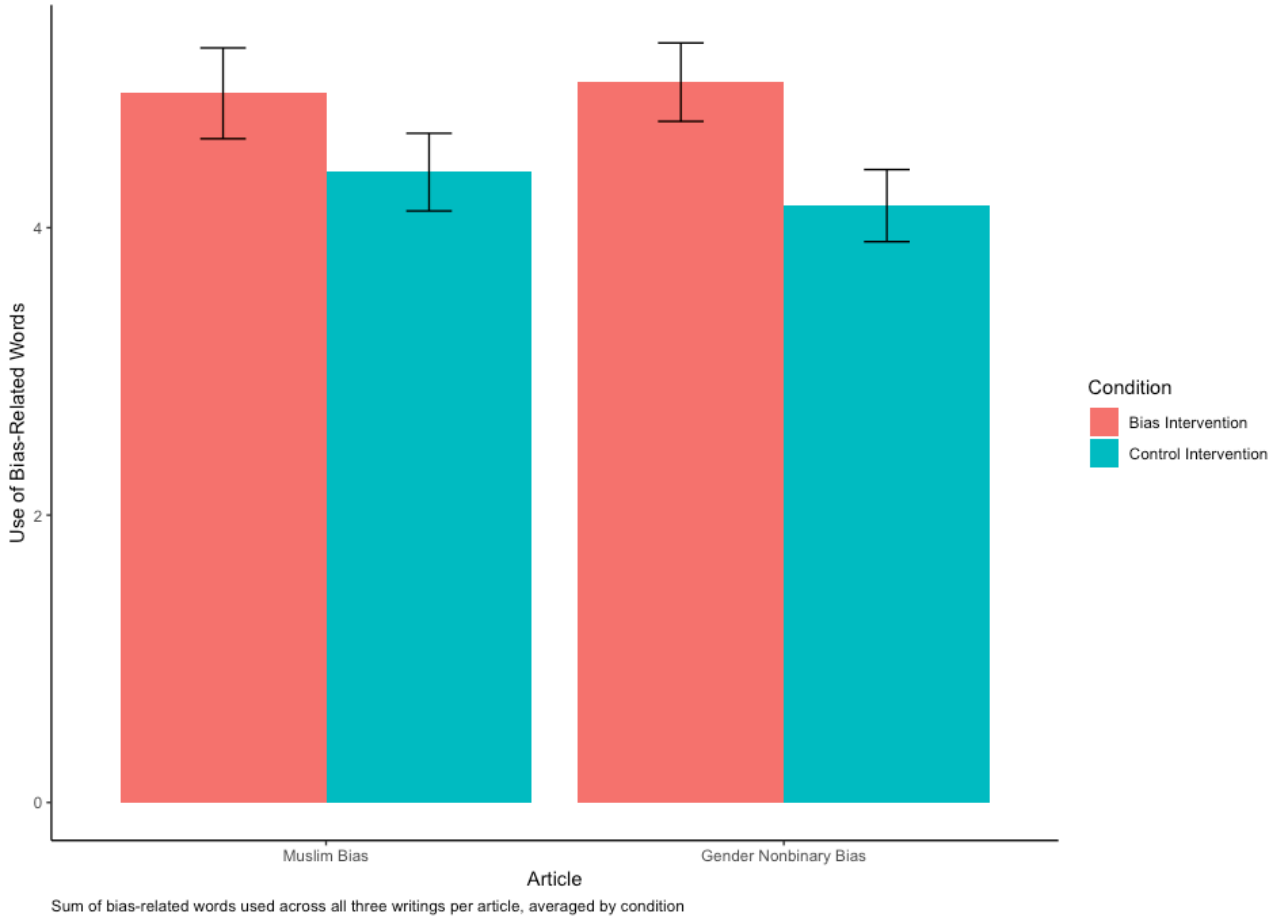

Figure 4. Sums of Bias/Inclusion Related Words Per Article by Condition

Finally, we were interested in whether the time between intervention and the Phase 2 study affected participants' likelihood of discussing bias/inclusion, as a way to further assess longevity of the training's effects. Time between Intervention and the Phase 2 study (1.25 - 2.15 years) did not directly affect the amount of bias-related words used (beta = -0.03, b = -0.04, p < 0.22), nor did it interact with Condition (beta = -0.03, b = -0.07, p = 0.26).

Discussion

To summarize, our Phase 2 findings provide compelling evidence showing the long-lasting, positive behavioral change created by Generation 2 of the bias habit-breaking

intervention. In general, bias training participants spoke up more about bias/inclusion in their writings compared to control participants. We think this pattern is especially compelling, given that this effect occurs despite Phase 2 occurring 1 to 2 years after the intervention was administered, and despite the fact that the training never mentioned the forms of bias/inclusion in these classroom discussions. Participants who received the training successfully generalized what they learned and applied them to new forms of bias/inclusion, beyond what they were explicitly taught.

Phase 3

Joining our past work with Generation 1 of the bias habit-breaking training approach (e.g., Devine et al., 2012; 2017; Forscher et al., 2017, see Cox & Devine, 2019 for a review), the present work validates Generation 2 of the bias habit-breaking training as an effective intervention that empowers people to reduce bias and create inclusion. The long term reduction in implicit bias observed in the initial randomized-controlled trial (Devine et al., 2012) was unprecedented in the literature at the time, and this study's replication of that effect in Phase 1 replicates that effect with a much larger sample size. In a recent meta-analysis of 492 experiments trying to reduce implicit bias (Forscher, Lai, et al., 2019), no method observed decreases in implicit bias that lasted more than 24 hours, with most lasting only a few minutes (see also Lai et al., 2016; Siden et al., 2021). Work on the bias habit-breaking training is the sole exception to these patterns. In fact, the bias habit-breaking training's approach is so substantially different from other interventions that Forscher, Lai, and colleagues excluded it from their meta-analysis.

The importance of a reduction in measured implicit bias lies in the assumption that this reduction will correspond to reductions in other outcomes; the IAT is very often used as a proxy for discriminatory judgments and behaviors (Cox & Devine, 2022). In the Forscher, Lai, et al. (2019) meta-analysis, when studies assessed implicit bias interventions alongside behavioral outcomes, observed decreases in IAT scores did not mediate corresponding reductions in biased behavior. This lack of mediation indicates that those reductions in measured implicit bias are unlikely to be meaningful for other outcomes (Forscher, Lai, et al., 2019). Theoretically, this lack of mediation makes sense, because most of the methods reviewed in that meta-analysis are incidental — participants complete some task (e.g., a priming task) at the behest of the experimenter, without understanding its purpose or engaging their conscious values related to bias reduction or regulation (Cox & Devine, 2022). In the bias habit-breaking training, however, participants are engaged as active agents in the change process and explicitly taught methods that they can use autonomously to reduce bias over time. This key difference in the model of change is theoretically crucial to the longevity of the implicit bias reduction observed by Devine and colleagues and the present study. Empowering participants as autonomous agents of change sets them up to continue to work on bias reduction over time. Further, we hypothesize that these effects should generalize to other measures related to bias and stereotyping. In Phase 3, we sought to assess the extent to which training participants were less likely to make stereotypic assumptions and evaluate the extent to which the reduction in implicit bias observed in Phase 1 would mediate stereotyping.

To the extent that the reduction in IAT bias observed in Phase 1 and in past work (Devine et al., 2012) reflects a meaningful reduction in bias, we would expect it to relate to other

measures relevant to bias reduction and bias regulation. We explore this in Phase 3, using Monteith and colleagues' (2002) stereotype regulation task (Burns et al., 2017; Czopp et al., 2006; Monteith et al., 2002). This task asks participants to make inferences about a set of stimulus people based on pictures and statements, and on key trials, it sets up participants with a prompt that strongly pulls for them to make a stereotypic inference. If they put in regulatory effort, however, they can avoid making the easy stereotypic response, to generate a non-stereotypic alternative. This task has been well-established as a measure of regulatory effort to avoid stereotyping, and we expect that bias training participants will show greater stereotype regulation (i.e., they will have fewer stereotype-congruent responses), compared to controls, and that this effect will be mediated by their measured reduction in IAT bias from Phase 1.

Method

Participants and Design. All participants from Phase 1 who completed at least 1 follow up survey were invited to participate via email. Similar to Phase 2, we sought to recruit 300 participants total. We successfully recruited 320 participants in total for Phase 3 (74.29% Female, 80.56% White, 44.51% Bias condition). Participants were compensated with a $20 Amazon Gift Card in exchange for their participation.

Procedure. We sent out emails to recruit participants, from a different email address than Phase 1 or 2, to minimize any possibility that the participants would connect the studies. The recruitment emails and consent described the study purpose as looking at interpersonal perceptions, studying how people can give descriptions about a person given limited information. Participants waived signed consent, and proceeded to complete the study on Qualtrics.

Stereotype Regulation Task. The stereotype regulation task were taken from stereotypic inferences task by Monteith and colleagues (2002; see also Burns et al., 2017; Czopp et al., 2006). In this task, participants are given a photograph and a brief description, and asked to make an inference about the person portrayed. Fourteen critical trials are constructed to heavily favor a stereotypic inference. For instance, one critical trial shows a picture of a Black man and the description "This person uses needles for recreation", which heavily draws for an inference that he is a needle drug user, matching stereotypes that Black men use drugs. If someone is putting effort into trying not to stereotype, however, they could provide an alternate response, for instance, that he likes knitting. Of the 14 critical trials, six involve stereotypes associated with Black men, and 8 involve stereotypes associated with White women. Thirty-two trials were designed to be filler unrelated to stereotypes. Participants typed their inferences in a text box.

Coding. All participants' responses to the gender and race questions were coded by two research assistants blind to participants' condition. Each response was independently coded by two research assistants blind to condition, to determine whether the response reflected stereotypes (1), or not (0). Inter rater reliability kappa was 0.813. A third coder resolved any coding difference produced by the first two coders. See Appendix B for the full coding scheme.

Results

Raw Counts. We first assessed the between-condition difference in the raw counts. When looking at the raw count data using an independent-samples t-test, bias training participants (M = 5.78, sd = 2.60) did not appear to stereotype less than environmental training participants (M = 6.02, sd = 2.36), t (318) = 0.846, p = 0.398.

Although there is no significant distal effect between Condition and the raw stereotyping counts, many recent scholars have argued that having a direct distal effect is not necessary to test mediation (Rucker et al., 2011, Shrout & Bolger, 2002; Zhao, Lynch, and Chen, 2010). The present data match the conditions set forth by Shrout and Bolger (2002) for one such case, dubbed the "suppression model", in which the bivariate distal effect obscures the complexity of the causal relations between the variables. We therefore proceeded to test whether Follow-up 2 IAT mediated the effect of Condition on the raw stereotyping counts.

The relationship between condition and stereotyping count was mediated by IAT score at FU2. As Figure 5 illustrates, the standardized regression coefficient between condition and IAT at FU2 was statistically significant, as was the standardized regression coefficient between FU2 IAT and stereotyping score. The standardized indirect effect was (-.13)(1.44) = -0.2. We tested the significance of this indirect effect using bootstrapping procedures. Unstandardized indirect effects were computed for each of 10,000 bootstrapped samples, and the 95% confidence interval was computed by determining the indirect effects at the 2.5th and 97.5th percentiles. The bootstrapped unstandardized indirect effect was -0.19, and the 95% confidence interval ranged from -0.43 to -0.02. Thus, the indirect effect was statistically significant.

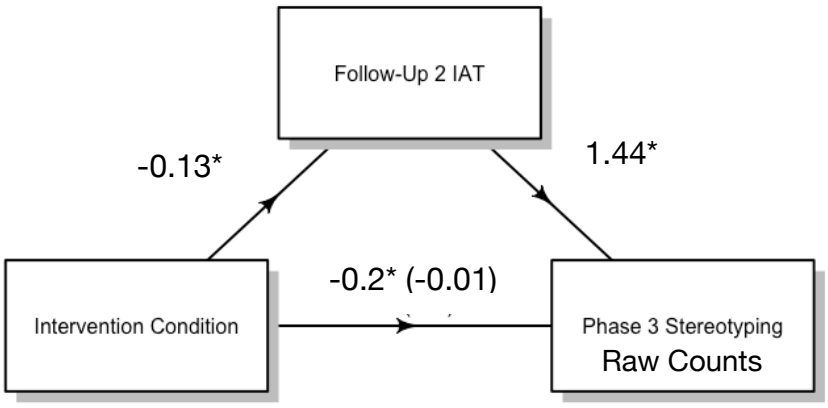

Figure 5. Mediation Model for Raw Count Stereotyping Rate

Binned Data. Although there was no significant mean-level difference when testing the raw counts, when looking at the *distributions* of the raw counts, the two conditions have very different distribution shapes (See Figure 6). Each distribution significantly violated normality assumptions (Shapiro-Wilk's > 0.95, *p*'s ≦ 0.002), with the bias training participants' distribution (Kurtosis = -0.883) being more strongly negatively kurtotic than the environmental training participants' distribution (Kurtosis = -0.355). Because of these violations of normality, we next binned the data into three categories so we could conduct a nonparametric analysis.

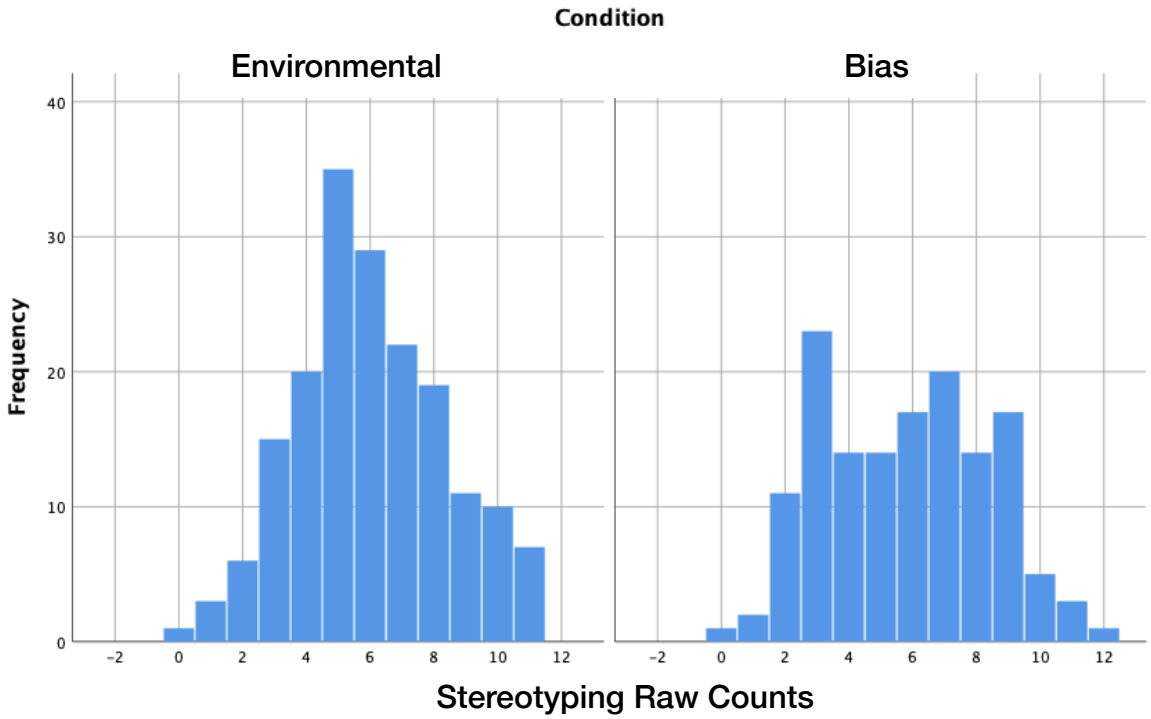

Figure 6. Raw Count Stereotyping Rate Distributions by Condition

There were 14 stereotype items total, but the most any participant stereotyped was 12 items, with zero being the least number of items any participant stereotyped. To conduct a chi-square, we grouped the stereotyping count into three levels of stereotyping: Low (participant

stereotyped on 0-3 items), Medium (participant stereotyped on 4-7 items), and High (participant stereotyped on 8-12 items). See Figure 7.

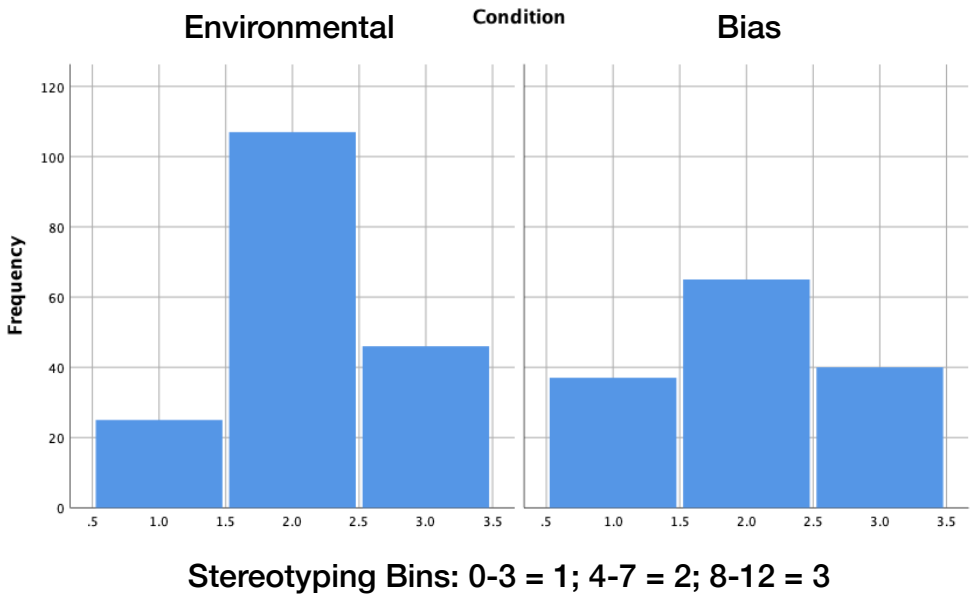

Figure 7. Binned Stereotyping Rate Distributions by Condition

A chi-square test of independence was performed to examine the relation between condition and the level of stereotyping. The nonparametric analysis on the binned stereotyping rates indicated that, matching hypotheses, bias participants were more likely to be in the low-stereotyping bin than control participants, X_2 (2, N = 320) = 8.93, p = 0.012. Bayes factor analyses (BF_{-0} Independent multinomial = 4.25) indicate that the data provide moderately strong evidence in favor of the hypothesis that Bias Intervention participants stereotyped less than Control participants.

Next, we tested the mediation model with the stereotyping rate bins. The relationship between condition and stereotyping level was in fact mediated by IAT score at FU2. See Figure 8. The standardized regression coefficient between condition and IAT was statistically significant (p = 0.0172), as was the standardized regression coefficient between IAT and stereotyping level

(p = 0.0005). We tested the significance of the indirect effect of the intervention on stereotyping via IAT using bootstrapping procedures. Unstandardized indirect effects were computed for each of 10,000 bootstrapped samples, and the 95% confidence interval was computed by determining the indirect effects at the 2.5th and 97.5th percentiles. The bootstrapped unstandardized indirect effect was -.04, and the 95% confidence interval ranged from -.08, -.01. Thus, the indirect effect was statistically significant.

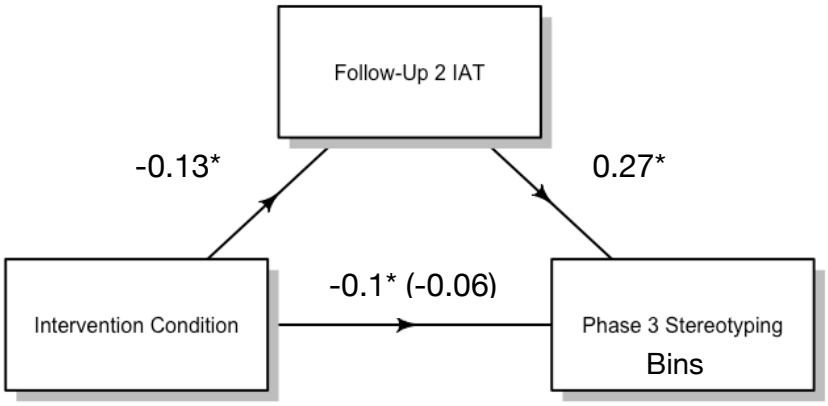

Figure 8. Mediation Model for Binned Stereotyping Rates

Discussion

Matching expectations, bias training participants were more likely than controls to put effort into generating non-stereotyping responses on the stereotype regulation task. Further, this effect was fully mediated by their level of IAT bias from 2 years earlier. We believe this pattern provides evidence that the training's observed long-lasting reduction in IAT bias is meaningful; this reduction is not an artifact of a given *measure*; it reflects a change in the overarching psychological *construct* of unintentional bias. Consistent with our theorizing and arguments in this and other articles, the bias habit-breaking training approach equips participants to recognize and overcome biases within themselves.

General Discussion

The abundant, increasingly publicized failures of diversity and bias trainings could lead people to feel discouraged, helpless, and defeated with regard to making positive changes related to bias, diversity, equity, and inclusion. In the present work, we add to the growing body of evidence that empowerment-based approaches to bias training can in fact be effective at creating lasting change. Whereas Generation 1 bias habit-breaking trainings focused solely on a single target group or context (i.e., anti-Black bias in Devine et al., 2012; Forscher et al., 2017; anti-woman bias in STEM in Carnes et al., 2015; Devine et al., 2017), the Generation 2 training developed and tested in the present article encouraged participants to address bias and inclusion issues related to any potential target group.

Phase 1 replicated past work (Devine, 1989) with a much larger sample size, such that bias training participants significantly decreased in IAT bias over time, whereas control participants did not. Further, in Phase 3, this decrease in implicit bias mediated performance on a stereotyping task 2-3 years later. Phase 2 provided evidence that training participants generalized what they had learned in the training to new forms of bias and inclusion, and were significantly more likely to speak up about bias and inclusion topics. These patterns are consistent with the theoretical expectations and practical intentions built into the bias habit-breaking training, that it should teach people generalizable, self-sustaining skills to regulate and reduce bias and promote inclusion.

Further evidence of the effectiveness of the habit-breaking training's general model of cognitive-behavioral change comes from the control condition of Phase 1, which adapted the same approach to the context of environmentalism and sustainability behaviors. Participants who

completed the environmental habit-breaking training engaged in more environmentally friendly, sustainable behaviors compared to control participants. Future work will further develop and more extensively test this approach.

The bias habit-breaking training is just one initial example of the benefit of adopting an empowerment-based approach (Cox, *under review*; Cox & Devine, 2019). It has been successful, where so many other trainings have failed, because 1) the training promotes active, self-sustaining change efforts, 2) it teaches customizable, generalizable tools that equip people to address many various forms of bias, 3) it is built on a solid, scientific model of cognitive-behavioral change, and 4) rather than trying to impose change on people, it respects their autonomy and empowers them to become agents of change themselves. In contrast to the considerable, widespread messages of hopelessness related to creating lasting change in diversity and bias contexts, empowerment-based approaches show considerable promise, and give reasons to hope that we can make positive, lasting changes.

References

al-Gharbi, M. (2020). Diversity-Related Training: What Is It Good For? *Heterodox: The Blog*. Retrieved from: https://heterodoxacademy.org/blog/diversity-related-training-what-is-it-good-for/ on 11/20/2021.

Allport, G. W. (1954). *The nature of prejudice*. Reading, MA: Addison-Wesley.

Amodio, D. M., Devine, P. G., & Harmon-Jones, E. (2007). A dynamic model of guilt: Implications for motivation and self-regulation in the context of prejudice. Psychological Science, 18, 524-530. Available from https://doi.org/10.1111/j.1467- 9280.2007.01933.x.

The Associated Press. (2017). UW-Madison Counted 74 Bias Incidents In Spring 2017. *Wisconsin Public Radio.*

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1), 1–48.

Birtel, M. D., & Crisp, R. J. (2015). Psychotherapy and social change: Utilizing principles of cognitive-behavioral therapy to help develop new prejudice-reduction interventions. *Frontiers in psychology*, *6*, 1771.

Bodenhausen, G. V., & Macrae, C. N. (1998). Stereotype activation and inhibition. *Advances in social cognition*, 11, 1-52.

Burns, M. D., Monteith, M. J., & Parker, L. R. (2017). Training away bias: The differential effects of counterstereotype training and self-regulation on stereotype activation and application. *Journal of Experimental Social Psychology*, 73, 97-110.

Carnes, M. L., Devine, P. G., Manwell, L. B., Byars-Winston, A., Fine, E., Ford, C. E., Forscher, P. S., Iaasc, C., Kaatz, A, Magua, W., Palta, M, & Sherridan, J. (2015). Effect of an intervention to break the gender bias habit: A cluster randomized, controlled trial. *Academic Medicine*, 90

Cox, W. T. L. (*Under review*). Developing scientifically validated bias and diversity trainings that work: Empowering agents of change to reduce bias, create inclusion, and promote equity

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

Cox, W. T. L. & Devine, P. G. (2019). The prejudice habit-breaking intervention: An empowerment-based confrontation approach. In Mallett, R. K., & Monteith, M. J. (Eds.). (2019). *Confronting prejudice and discrimination: The science of changing minds and behaviors.* Academic Press., London, UK. 249–274.

Cox, W. T. L. & Devine, P. G. (2022). Changing implicit bias vs Empowering people to address the personal dilemma of unintentional bias. *In NSF book on Implicit Bias*.

Cox, W. T. L., Abramson, L. Y., Devine, P. G., & Hollon, S. D. (2012). Stereotypes, prejudice, and depression: The integrated perspective. *Perspectives on Psychological Science*, *7*(5), 427-449.

Czopp, A. M., Monteith, M. J., & Mark, A. Y. (2006). Standing up for a change: Reducing bias through interpersonal confrontation. *Journal of Personality and Social Psychology*, 90, 784–803.

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

Devine, P. G., & Ash, T. L. (2021). Diversity Training Goals, Limitations, and Promise: A Review of the Multidisciplinary Literature. *Annual review of psychology*, *73*.

Devine, P. G., & Monteith, M. J. (1993). The role of discrepancy associated affect in prejudice reduction. In D. M. Mackie, & D. L. Hamilton (Eds.), *Affect, cognition, and stereotyping* (pp. 317-344). New York: Academic Press.

Devine, P. G., Forscher, P. S., Austin, A. T., & Cox, W. T. L. (2012). Long-term reduction in implicit race bias: A prejudice habit-breaking intervention. *Journal of Experimental Social Psychology, 4*.

Devine, P.G., Forscher, P.S., Cox, W. T. L., Sherridan, J. Kaatz, A., Carnes, M.L. (2017). A gender habit-breaking intervention led to increased hiring of female faculty in STEMM departments. *Journal of Experimental Social Psychology, 73*, 211-215.

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.

Dix, E. L., Harris, B. M., & Devine, P.G. (under review). White People's Receptivity to Black People's Confrontations of Bias: Concern About Discrimination.

Fiske, S. T. (1998). Stereotyping, prejudice, and discrimination. In D. T. Gilbert, S. T. Fiske, & L. Gardner (Eds.), *The handbook of social psychology* (pp. 357–411). New York, NY: McGraw-Hill.

Forscher P. S. (2017). The individually-targeted habit-breaking intervention and group-level change. Thesiscommons.org/4t7fy

Forscher, P. S., Lai, C. K., Axt, J. R., Ebersole, C. R., Herman, M., Devine, P. G., & Nosek, B. A. (2019). A meta-analysis of procedures to change implicit measures. *Journal of personality and social psychology*, 117(3), 522.

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

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

Greenwald, A. G., Nosek, B. A., & Banaji, M. R. (2003). 'Understanding and using the Implicit Association Test: I. an improved scoring algorithm': Correction to Greenwald et al. (2003). *Journal of Personality and Social Psychology*, 85, 481. Available from https://doi.org/10.1037/h0087889.

Hunt (2016). How Green is Your Life? http://www.sustainablestuff.co.uk/quiz-how-green-your-life.html

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(8), 1001–1016.

Lee, M. D., & Wagenmakers, E. J. (2014). *Bayesian cognitive modeling: A practical course.* Cambridge university press.

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. Available from https://doi.org/10.1037/0022-3514.65.3.469.

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*(4), 901.

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.

Moss-Racusin, C. A., van der Toorn, J., Dovidio, J. F., Brescoll, V. L., Graham, M. J., & Handelsman, J. (2014). Scientific diversity interventions. *Science*, 343(6171), 615-616.

Paluck, E. L. (2006). Diversity training and intergroup contact: A call to action research. *Journal of Social Issues*, 623, 439!451. Available from https://doi.org/10.1111/j.1540-4560.2006.00474.x.

Paluck, E.L., Porat, R, Clark, C. S., Green, D. P. (2021). Prejudice reduction: progress and challenges. *Annual Review of Psychology*. 72:14.1-14.28

Pendry, L. F., Driscoll, D. M., & Field, S. C. (2007). Diversity training: Putting theory into practice. *Journal of Occupational and Organizational Psychology*, 80, 27!50. Available from https://doi.org/10.1348/096317906X118397.

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. (2009). The active control of prejudice: Unpacking the intentions guiding control efforts. *Journal of Personality and Social Psychology*, 96, 640-652.

Prochaska, J. O., & Velicer, W. F. (1997). The transtheoretical model of health behavior change. *American journal of health promotion*, *12*(1), 38-48.

Rucker, D. D., Preacher, K. J., Tormala, Z. L., Petty, R. E. (2011). Mediation analysis in social psychology: Current practices and new recommendations. *Social and Personality Psychology Compass, 5*(6), 359-371.

Shrout, P. E., & Bolger, N. (2002). Mediation in experimental and nonexperimental studies: New procedures and recommendations. *Psychological Methods, 7*(4), 422–445. https://doi.org/10.1037/1082-989X.7.4.422

Siden, J. Y., Carver, A. R., Meja, O. O., Townsel, C. D. (2021). Reducing implicit bias in maternity care: A framework for action. *Women's Health Issues.*

Stevenson, S. (2017). Fighting racial bias on campus. *The New York Times*.

Zhao, X., Lynch Jr, J. G., & Chen, Q. (2010). Reconsidering Baron and Kenny: Myths and truths about mediation analysis. *Journal of consumer research*, *37*(2), 197-206.