Running Head: THE LIMITED MALLEABILITY OF IMPLICIT EVALUATIONS

The Surprisingly Limited Malleability of Implicit Racial Evaluations

Jennifer A. Joy-Gaba

Brian A. Nosek

University of Virginia

Word count: 5,463

Author's Note: Correspondence concerning this article should be sent to Jennifer Joy-Gaba, Department of Psychology, University of Virginia, P.O. Box 400400, Charlottesville, VA 22904. Electronic mail may be sent to jaj3f@virginia.edu.

Abstract

Implicit preferences for Whites compared to Blacks can be reduced via exposure to admired Black and disliked White individuals (Dasgupta & Greenwald, 2001). In four studies (total *N*=4,628), while attempting clarify the mechanism, we found that implicit preferences for Whites were weaker in the "positive Blacks" exposure condition compared to a control condition (weighted average *d*=.08). This effect was substantially smaller than the original demonstration (Dasgupta & Greenwald, 2001; *d*=.82). Factors beyond exposure to admired Blacks may be necessary for the effect, such as making race accessible during exemplar exposure, and including negative White exemplars. Our evidence suggests that exposure to known-group members shifts implicit race bias reliably, but weakly.

Words=110

Keywords: malleability, implicit attitudes, race bias, Implicit Association Test, social cognition

The Surprisingly Limited Malleability of Implicit Racial Evaluations

Schneider and Schiffrin (1977) introduced automaticity as processes that can be activated without control, requiring little to no conscious awareness to initiate or complete. Automatic processing is now understood to be a key component of all aspects of human behavior (Bargh & Chartrand, 1999). Early notions of automaticity suggested that the processes were difficult to modify because of the well-learned patterns that were relatively insensitive to the immediate context. This special issue adds to the growing body of evidence challenging the assumption of automatic inflexibility (Blair, 2002; Dasgupta & Asgari, 2004; Mitchell, Nosek, & Banaji, 2003; Sinclair, Lowery, Hardin, & Colangelo, 2005). Indeed, social cognition research introduced a new understanding of automaticity as contextually sensitive and amenable to change.

A large portion of the research on malleability of automatic processes focuses on implicit social group biases such as associating positive concepts with White people more easily than with Black people. Such associations are pervasive and can exist even in those who espouse egalitarian values (Nosek, Smyth, et al., 2007). As such, there are personal and social factors that may elicit shifts in the activation or expression of implicit racial biases. Multiple interventions effectively shift implicit racial biases, at least temporarily (e.g., Dasgupta & Greenwald, 2001; Lowery, Hardin & Sinclair, 2001; Sinclair, et al., 2005; Wittenbrink, Judd, & Park, 2001). Dasgupta and Greenwald (2001, hereafter "DG"), for example, exposed participants to 10 admired Black (e.g., Martin Luther King, Jr.) and 10 disliked White individuals (e.g., Charles Manson) before completing an Implicit Association Test (IAT; Greenwald, McGhee, & Schwartz, 1998). Participants viewed each exemplar four times. During the first two viewings, participants

had to decide which of two short descriptions accurately described the person. During the last two viewings, participants categorized each person as Black or White. Compared to a control group, participants viewing admired Black and disliked White individuals showed less implicit preference for White people compared to Black people (d=.82[1]). The effect persisted in a follow-up test 24 hours later (d=.71). In a conceptual replication, similar reductions in implicit pro-young bias were observed after viewing admired elderly and disliked young individuals bias compared to viewing disliked elderly and admired young individuals (d=.89).

DG's demonstration is a heavily cited example of the factors that can shift implicit racial biases. A non-random sampling of articles suggests that one popular interpretation of DG's effect is that exposure to admired Black individuals is the key causal influence. For example, Foroni and Mayr (2005) suggest the malleability effect comes from "exposing participants to examples of positively viewed blacks" (p. 139). Rudman, Ashmore, and Gary (2001) suggested that individuals "exposed to positive Black exemplars" show less implicit Pro-White bias (p. 858). And, Mitchell, Nosek, and Banaji (2003, p. 456) concluded that "exposure to positive African American exemplars resulted in participants producing evaluations of that group that were not as negative as those produced in a control condition." DG themselves were more cautious in asserting that the effect resulted from exposure to both admired Black and disliked White individuals.

We initiated the present research to examine mechanisms underlying the malleability of social evaluations, with particular interest in DG's paradigm. Like the authors above, we started with the assumption that the critical component of the

[1] The effect sizes are provided by Dasgupta and Greenwald through correspondence.

manipulation was exposure to admired Black individuals and implemented our manipulation accordingly (Experiment 1). Despite a very large sample (*N*=1,303), we did not replicate the original malleability effect. Our goal then shifted to replicating the original effect. In so doing, we hoped to shed light on the factors contributing to the malleability of implicit social evaluations. In Experiments 2a (*N*=944) and 2b (*N*=1,191), after adding negative White exemplars and requiring categorization of the exemplars by race, we successfully replicated the effect, though with a relatively weak effect magnitude (*d*=.17 and *d*=.14 respectively, compared to *d*=.82 in DG). In Experiment 3 (*N*=1,190), we tested whether unique features of our setting or sample could explain the weaker malleability effects. They could not. Together, these studies suggest that exposure to known-group members can reliably shift implicit race bias as measured by the IAT, but that this effect may be relatively weak.

Experiment 1

The goal for Experiment 1 was to compare two possible interpretations of the malleability effect in the DG paradigm. One possibility is that exposure to positive Black individuals temporarily strengthens the association between *Black people* and *good* (or weakens the association between *Black people* and *bad*). Alternatively, exposure to admired Black individuals might activate goals to be egalitarian that inhibit the expression of implicit race bias (Moskowitz, et al., 1999; Devine, et al., 2002). In this case, we hypothesized that a goal to be egalitarian would reduce implicit racial biases *and* implicit biases for other stigmatized groups (e.g., old people). Because we did not replicate the original malleability effect, this report does not dwell on these hypotheses. Instead, we focus on analyses relevant to the original malleability effect.

Participants viewed exemplars from one of four categories: admired elderly individuals, admired Black individuals, admired female scientists, or cartoon characters (control). Participants completed four IATs: age attitudes, race attitudes, season attitudes (control), and stereotypic associations between gender and science versus humanities.

Method

Participants

1,403 students from Brock University participated as part of a course requirement. We included only those participants who had accurately identified six out of the eight exemplars (75% correct) leaving 1,303 students (965 female, 333 male, 5 unreported, mean age=19.8 years)[2]. Of these, 1,060 (82.2%) were White, and the rest were from another racial category. Participants completed the experiment on the Project Implicit research infrastructure (http://projectimplicit.net/) via a private link.

A power analysis for the key significance test revealed 62% power for detecting a small Cohen's *d* effect of .15, and 99% power to detect a Cohen's *d* of greater than .30. Using DG's original effect size of .82 as a benchmark, the sample size virtually guaranteed that the effect would be observed.

Materials

Positive Exemplar Induction. Participants viewed eight well-known individuals from one of three categories: admired elderly individuals, admired Black individuals, or admired female scientists. Although some exemplars' group memberships overlap between categories (e.g., Nelson Mandela is both Black and elderly), within each category only one of the attributes represented all members of that condition. A control

[2] Results from the current experiment and the additional experiments remain unchanged when this criterion is not imposed.

condition presented eight popular cartoon characters. In the induction phase, each exemplar's name and image appeared above one correct and one incorrect description of the person's achievements. Both statements were positive so that if the person did not know the individual, they would form a positive impression whichever description they selected. In the control condition, the statements were neutral descriptions of the cartoon characters.

The induction procedure required that participants select the accurate description for each individual. Half of the correct descriptions appeared first. A complete list of the stimuli for this experiment and the other experiments are available in supplementary materials at http://briannosek.com.

Implicit Association Test (IAT; Greenwald et al., 1998). The IAT measures the strength with which concepts (e.g., Black and White people) are associated with attributes (e.g., good and bad). Stimulus items representing four categories are categorized one at a time and as quickly as possible using two keys of a computer keyboard. During the first critical response block *Black faces* and *good words*, for example, are categorized with the same response key, while *White faces* and *bad words* are categorized with a second response key. In the second critical response block, the response pairings are reversed such that *Black faces* and *bad words* are categorized with one key, and *White faces* and *good words* with the other key. If the concepts sharing a response key are associated, it should be easier to categorize the stimulus items quickly compared to when the concepts sharing a response key are not associated (see Nosek, Greenwald, & Banaji, 2007 for a review). Participants completed four IATs measuring age, race, and season associations with evaluation (good and bad), and a measure of

associations between gender and academic domains (science and humanities). The specific IAT procedure followed recommendations of Nosek, Greenwald, and Banaji (2005), and data analysis used the *D* algorithm (Greenwald, Nosek, & Banaji, 2003).

Self-Report Measures. Participants self-reported feelings toward Black people and White people, old people and young people, summer and winter seasons. Given that the results revealed no explicit malleability for any of the presented experiments and given our framing of this article, we only report the implicit measure results.[3]

Procedure

Participants completed one of the four exposure inductions manipulated between-subjects. Then, participants completed all four IATs and all self-report measures. The presentation order of the measures was randomized with the caveat that all the measures of one type (implicit or explicit) were completed together.

Results

Descriptive Statistics

Replicating previous research (Nosek, Smyth, et al., 2007; Nosek & Hansen, 2008), participants implicitly preferred White people to Black people (M=.32, SD=.35), $t(1283)=33.26$, $p<.01$, d=.93 , younger people to older people (M=.31, SD=.36), $t(1298)=30.86$, $p<.01$, d=.86, and summer to winter (M=.53, SD=.41), $t(1265)=46.79$, $p<.01$, d=1.32. In addition, participants implicitly associated males with science and females with liberal arts more than the reverse (M=.26, SD=.35), $t(1277)=26.40$, $p<.01$, d=.74.

[3] Across studies, explicit attitude reports were positively related to the corresponding IAT scores (Experiment 1: race, r=.21, $p<.01$, age, r=.14, $p<.01$, gender, r=.22, $p<.01$, seasons r=.29, $p<.01$, Experiment 2a: r=.23 , $p<.01$, Experiment 2b: r=.18, $p<.01$, Experiment 3: Project Implicit volunteers, r=.20, $p<.01$, online undergraduates r=.20, p=.05, lab undergraduates r=.28, $p<.01$).

Searching for Implicit Malleability

If exposure to positive exemplars of a category strengthened the association between that category and good (or weakened the association with bad), then participants should have less implicit pro-White bias on the race IAT when primed with admired Black individuals, less implicit pro-young bias when primed with admired elderly individuals, and less implicit gender stereotypes when primed with admired female scientists, compared to the other conditions. This did not occur. As shown in Figure 1, a 4 (type of IAT) x 4 (condition) mixed effects ANOVA revealed no significant differences between exposure conditions, $F(1, 1251)=.25$, $p=.62$, $d=.01$, and no interaction between condition and the type of IAT, $F(3, 3753)=.59$, $p=.62$, $d=.01$. There was a significant difference in effect magnitudes between IATs, $F(3, 3753)=35.96$, $p<.01$, $d=.10$ showing that some IAT effects (summer-winter) were larger than others (gender stereotypes), but this is not relevant to our predictions.

Given our 4x4 design and focused hypothesis, the ANOVA is a very low-powered test. Contrast coding is more appropriate. We conducted two contrasts for each of the three exposure conditions (e.g., admired Blacks, admired elderly people, admired female scientists) compared to the control exposure condition (cartoon characters). For one contrast, we compared target exposure condition (e.g., admired Blacks) to the other three exposure conditions (e.g., age, gender, control) on the target IAT (race). This contrast was not reliable for race ($p=.22$, $d=.07$), age ($p=.89$, $d=.01$), or gender-science ($p=.15$, $d=.08$). The other contrasts compared the three "admired people" exposure conditions (e.g., race, age, gender) versus the control condition on each IAT. Again, none of these

contrasts were reliable (race IAT: p=.31, d=.06; age IAT: p=.33, d=.05; gender IAT: p=.11, d=.09).

Focused analyses on race

Because of the age and gender exposure conditions, none of the preceding analyses are identical to DG's *t*-test comparing a race exposure condition to a control condition. We isolated the race and control conditions. A *t*-test revealed that viewing positive Black exemplars did not reduce implicit pro-White bias compared to viewing cartoons, $t(670)$=.09, p=.93, d=.01. It is also possible that the inclusion of four IATs in a randomized order weakened the manipulation for the IATs completed at the end. When we compared the race exposure and control conditions for only those that had the race IAT first, there still was no condition difference, $t(173)$=.12, p=.90, d=.02.

Age. DG also showed a malleability effect on an age attitude IAT after presenting positive elderly and negative young exemplars. We did not replicate this effect either. A *t*-test revealed that viewing positive elderly exemplars did not reduce implicit pro-young bias more than viewing cartoons, $t(673)$=.52, p=.61, d=.04. When we examined just the participants that completed the age IAT first within the implicit measures there still was no effect, though the effect size appeared larger, $t(143)$=1.09, p=.28, d=.18.

Gender. Dasgupta and Asgari (2004) showed that viewing admirable female leaders (with no negative males for contrast), reduced implicit stereotypes about gender and leadership versus supportive roles for female participants. Our conceptually related exposure to female scientists did not reduce implicit gender-academic stereotypes compared to the control condition, $t(541)$=1.83, p=.07, d=.16. And, looking at just the

participants completing the gender-academic IAT first did not reveal a significant malleability effect, $t(117)=.63$, $p=.53$, $d=.12$.

Discussion

Despite a very high-powered design, we were unable to replicate the reduction of implicit race (or other) bias after exposure to admired members of the stigmatized category. There were at least two procedural differences between the current experiment and DG that, in retrospect, might be important for the malleability effect. In DG, participants viewed each exemplar four times: twice while matching the individual's name and picture with a description of why the person is admired, and twice to categorize each exemplar by name as a Black or White individual. In contrast, this experiment presented each admired person just once, and the category of interest (e.g., race) was never highlighted. It is possible that repeated exposure and making race accessible are important features for obtaining the malleability effect.

Also, as already anticipated, DG's included negative exemplars of the dominant group during the exposure, but the present study only used positive exemplar exposure. To increase our likelihood of replicating the original effect, we narrowed the goal of Experiment 2 to replicating the race malleability effect with a paradigm more similar to DG.[4]

Experiment 2a and 2b

We increased the similarity of our manipulation to DG by making three changes to Experiment 1 materials: (1) added 10 disliked Whites, (2) added 2 exemplars for

[4] There are other slight differences between the paradigms that we did not test in this article such as: (a) DG provided feedback on participants' accuracy during the induction phase, but we did not, (b) some of the exemplars and descriptions used in the induction phase were different, and (c) we used faces as race exemplars in the IAT whereas DG used names.

admired Blacks (now totaling 10), and (3) added a race categorization task. After the first round of viewing each exemplar and matching with their admired or despised qualities, a second phase required participants to categorize each exemplar's name with their racial identity.

We added a third exposure condition comprised of admired Black and *admired* White individuals. We intended this to continue our initial goal of clarifying whether the effect reflected a temporary shift in association strengths or a more general goal to be egalitarian by seeing admired people of multiple ethnicities. Given the framing of this article, we do not dwell on this condition.

Experiment 2b is a direct replication of Experiment 2a, so they are described together.

Method

Participants

Experiment 2a. 951 individuals visiting the Project Implicit research website (https://implicit.harvard.edu/) were randomly assigned to this study from a pool of dozens of studies. We included only those participants who had accurately identified 30 out of the 40 exemplars (75% correct) during the induction procedure leaving 944 individuals (615 female, 329 male, mean age=28.8 years) for analysis. Once assigned to this study, participants were never again assigned to it on subsequent visits. 796 (75.2%) participants were White, and the rest were from another racial category. African Americans were excluded from this experiment because on average, they show little implicit preference between Whites and Blacks on the IAT (Nosek, Smyth, et al., 2007).

This design had 58% power in detecting a small Cohen's *d* effect of .15, and 98% power to detect a Cohen's *d* of greater than .30 for the critical replication test.

Experiment 2b. 1,200 individuals visiting the Project Implicit research website were randomly assigned to this study. Our analysis included only those participants who accurately identified 30 out of the 40 exemplars (75% correct) leaving 1,191 participants (752 female, 432 male, 7 unreported, mean age=19.8 years) for analysis. 1022 (88.3%) participants were White, and the rest were from another racial category except for African American.

This design had 69% power in detecting a small Cohen's *d* effect of .15 for the key analysis, and greater than 99% power to detect a Cohen's *d* of greater than .30.

Materials

Category Primes. We selected well-known individuals for each of three categories: 10 admired Black individuals, 10 admired White individuals, or 10 disliked White individuals. Each category was comprised of 2 women and 8 men. The categories were combined such that the 10 admired Black individuals were either paired with 10 admired White individuals or 10 disliked White individuals. Pictures of 10 flowers and 10 insects served as stimuli for the control condition. A complete list of the stimuli appears in the supplement.

Implicit Association Test (IAT; Greenwald et al., 1998). The racial attitude IAT followed the same procedure as described in Experiment 1.

Self-Report Measures. Participants completed explicit measures of racial attitudes toward Blacks and Whites as described in Experiment 1.

Procedure

The induction had two phases. First, participants selected which of two descriptions accurately described each of the 10 exemplars. Second, participants viewed the name of each exemplar again and categorized each person as Black or White (experimental conditions) or each object as a flower or insect (control condition). The induction phase was followed by the implicit and explicit measures presented in a randomized order.

Results

Experiment 2a

Overall, participants implicitly preferred Whites to Blacks (*M*=.40, *SD*=.37), *t*(931)=32.79, *p*<.01, *d*=1.07. Comparing the two key conditions, a *t*-test revealed that the implicit preference for Whites compared to Black was significantly weaker after exposure to admired Blacks and despised Whites (*M*=.36, *SD*=.36) compared with exposure to flowers and insects, (*M*=.42, *SD*=.38), *t*(589)=2.08, *p*=.04, *d*=.17.[5] The third condition, exposure to admired Blacks and admired Whites (*M*=.41, *SD*=.37), elicited an implicit preference for Whites over Blacks very similar to the control condition, *t*(598)=-.36, *p*=.72, *d*=-.03, and only marginally larger than the admired Black and disliked White exposure condition, *t*(642)=1.79, *p*=.07, *d*=.14.

Experiment 2b

Participants implicitly preferred Whites to Blacks (*M*=.35, *SD*=.41), *t*(1172)=29.27, *p*<.01, *d*=85. A *t*-test revealed that the implicit preference for Whites compared to Black was significantly smaller after exposure to admired Blacks and

[5] A 2 X 2 ANOVA with the order of implicit and explicit measures added as a factor elicited no significant interaction, *F*(2, 929)=.31, *p* =.73, *d*=.04, suggesting that this effect did not differ whether the IAT was completely before or after self-report measures. The same was observed for Experiment 2b, *F*(2, 1170)=.16, *p* =.85, *d*=.02, and Experiment 3, *F*(2, 1177)=.78, *p* =.46, *d*=.05.

despised Whites (*M*=.32, *SD*=.41) compared with exposure to flowers and insects, (*M*=.37, *SD*=.42), *t*(788)=1.99, *p*=.05, *d*=.14. The third condition, exposure to admired Blacks and admired Whites (*M*=.36, *SD*=.40), elicited effects similar to the control condition, *t*(738)=-31, *p*=.76, *d*=.02, and only marginally larger than the admired Black and disliked White exposure condition, *t*(775)=1.69, *p*=.09, *d*=.12.

Discussion

After adding negative White exemplars and making race accessible with a racial categorization task, Experiment 2a and 2b replicated DG's original finding that viewing admired Black and disliked White individuals elicits a weaker implicit race bias on the IAT compared to a control condition. However, our effect was much weaker (>70% smaller) than demonstrated by DG. An effect size less than .20 may have different practical significance in designing interventions aimed at reducing implicit race bias than an effect size of greater than .80. As such, it is important to understand the extent to which implicit social biases can be changed as a function of exposure to members of social groups. Also, it is notable that viewing admired Blacks and Whites did not elicit effects any different than the control condition suggesting that the exposure does not elicit a general egalitarian motivation, but instead shifts the accessible associations of Whites and Blacks with positivity and negativity. Given that Experiments 2a and 2b showed an effect that was not observed in Experiment 1 suggests that exposure to negative members of the dominant group and making race accessible may be important contributors to the effect. We return to this in the General Discussion.

It is possible that our large samples provide a more accurate estimate of the actual malleability effect size in this paradigm. Alternatively, there may still be important

differences between our sample and procedure that account for the different effect magnitudes.

DG used a 'typical' sampling method – participant pool undergraduates in a University laboratory. Our first three studies were conducted via the Internet, a difference in situation, and two of the studies (2a and 2b) used a heterogeneous volunteer sample, a difference in sample. In the context of the discrepant results, these sampling differences introduce the possibility that (a) malleability effects are more difficult to obtain via the Internet than in the laboratory, and (b) heterogeneous volunteer samples are less susceptible to malleability effects than student samples.

Internet versus laboratory. A simplistic explanation such as "the Internet does not work for experimental manipulations" is false. The Internet is now well-established as an effective tool for experimental research (Skitka & Sargis, 2005). In particular, the Project Implicit Virtual Laboratory has been used effectively for similar investigations (e.g., Bar-Anan, Nosek, & Vianello, 2009; Graham, Haidt, & Nosek, 2009; Lindner & Nosek, 2009; Nosek, 2005; Nosek & Hansen, 2008; Nosek, Smyth, et al., 2007; Ranganath & Nosek, 2008). However, a potentially important difference between the Internet and laboratory in this context is the presence of an experimenter. An experimenter's presence might amplify malleability effects if the participant infers the experimenter's beliefs from the exemplar exposure. This hypothesis derives from social tuning research; in some social circumstances, individuals automatically shift their social evaluations to be consistent with the social expectations of others (Sinclair, et al., 2005). Although DG did not have the experimenter explicitly endorse egalitarian ideals, highlighting admired Black and disliked White individuals may have led participants to

interpret the social values of the experimenter. As a result, participants may have "tuned" to the implied egalitarian values amplifying the effect of viewing admired Blacks and disliked Whites.

We conducted an experiment (N=602) attempting to increase the "presence" of the experimenter on the Internet by presenting a video of the experimenter giving the instructions making clear that they were completing *her* experiment. This manipulation did not increase the impact of the exemplar exposure compared to a no video condition, $t(525)=.52$, $p=.60$. The manipulation may not have been strong enough to elicit social tuning, so in Experiment 3 we directly replicated Experiments 2a and 2b, but manipulated the social circumstance of data collection: laboratory versus Internet.

Students versus volunteer heterogeneous sample. The other obvious sampling distinction between Experiments 2 and 3 and the original demonstration is our use of a heterogeneous volunteer pool. While there is no obvious reason why, it is possible that volunteers, despite being randomly assigned to this study, are importantly different than "typical" undergraduate samples for eliciting malleability. Lending some doubt to this possibility, Experiment 1 used an exclusively Introductory Psychology sample and found no malleability effect. Even so, in Experiment 3 we compared undergraduate and heterogeneous volunteer samples.

Experiment 3

Experiment 3 was a direct replication of Experiments 2a and 2b adding an orthogonal factor that manipulated setting (Internet or laboratory) and sample (undergraduate participant pool or heterogeneous volunteers). Students in the participant pool at the University of Virginia completed the study either online or in the laboratory.

Once the undergraduate participants signed-up to complete the experiment, they were randomly assigned to condition. A third condition was identical to the previous two studies – web volunteers. All participants completed experiment on the Project Implicit research infrastructure (http://projectimplicit.net/) using the identical study materials.

Method

Participants

Internet volunteers. Of the 1,027 individuals completing the study via the Project Implicit research website, 1,018 participants (704 female, 310 male, 4 unreported) met our manipulation cut-off by correctly identifying 30 out of the 40 exemplars (75% correct) during the induction procedure. 775 (75.2%) participants were White, and the rest were from another racial category.

Undergraduate sample. 174 undergraduates from the University of Virginia completed the study for partial course credit. 172 answered at least 30 of the 40 (75%) manipulations correctly during the induction procedure. Of those, 95 (71 female, 24 male) completed the study via the Internet. The remaining 77 (55 female, 21 male, 1 unreported) completed the study in the laboratory. Both undergraduate lab and web samples had greater than 80% power (81% and 88% respectively) to detect a malleability effect of the same magnitude as DG.

Materials and Procedure

The materials and procedure were identical to Experiment 2a and 2b except that the condition with admired Black and admired White exemplars was removed leaving two conditions [1] exposure to admired Black and disliked White exemplars, and [2] exposure to flowers and insects.

Results and Discussion

As shown in Figure 2, a 2 (condition) x 3 (sample) ANOVA revealed no main effect of condition, $F(1,1178)=.76$, $p=.74$, $d=.05$, and no interaction between condition and sample, $F(2,1177)=.04$, $p=.96$, $d=.01$. Despite being a direct replication of the two key conditions from the previous studies, this result is a failure to replicate the (weak) malleability effect.

Considering each sample individually, among the web volunteers, the exposure to admired Black and disliked White individuals ($M=.30$, $SD=.42$) did not elicit significantly weaker implicit race bias compared to viewing flowers and insects ($M=.32$, $SD=.38$), $t(999)=.96$, $p=.34$, $d=.03$. Among the undergraduate online sample, there was no difference between the race exposure condition ($M=.44$, $SD=.33$) and the control condition ($M=.47$, $SD=.29$), $t(93)=.51$, $p=.61$, $d=.11$ though the effect size was not far from our previous demonstrations. Finally, for the undergraduate lab sample, there was no difference between the race exposure condition ($M=.30$, $SD=.32$) and the control condition ($M=.34$, $SD=.29$), $t(75)=.70$, $p=.48$, $d=.16$. Again, however, the effect size was similar to the previous two experiments. Many more participants would be needed to estimate the effect reliably. Lab-based undergraduates showed weaker implicit pro-White bias compared to web-based undergraduates, $F(1, 170)=8.44$, $p<.01$, $d=.45$. This was not hypothesized a priori, but could reflect the presence of greater motivation to overcome bias when in a social situation with others (the laboratory) compared to being on the Internet (Evans, et al., 2003; Devine et al., 2002). Web volunteers ($M=.30$, $SD=.42$) however, showed effects similar to the laboratory participants, although the magnitude of the effect was smaller. In any case, even if different motivations were at

play between conditions, they did not alter the effectiveness of the exposure manipulation.

While Experiment 3 failed to replicate the malleability effect, it is notable that the effect size for the undergraduate samples was similar to the malleability effect magnitude observed in Experiments 2a and 2b (*d*'s~.15). Importantly, the study materials were identical to the successful demonstrations in the previous two studies suggesting that the failure to replicate is indicative of the weak malleability effect rather than due to a change in materials. This makes it difficult to conclude that the weak malleability effect is a consequence of sampling heterogeneous web volunteers or using the Internet as the mechanism for data collection. When the identical study materials were given to student samples in the lab or online, similarly small effect sizes were obtained.

General Discussion

We examined the malleability of implicit social preferences in response to exposure to admired and disliked group members. Some features of the exposure paradigm – such as presenting negative exemplars of the preferred group and making the social category (e.g., race) accessible – appear to be important for eliciting malleability. Further, while DG reported a large effect of exposure on implicit racial (and age) preferences (*d*=.82), the effect sizes in our studies were considerably smaller. None exceeded *d*=.20, and a weighted average by sample size suggests an average effect size of *d*=.08, or *d*=.10 for just Experiments 2a, 2b, and 3 – the replicable paradigm. All told, these results suggest that (a) exposure to positive and negative exemplars of social categories is a replicable method for shifting implicit social biases (particularly focused on race in these demonstrations), (b) the degree of malleability may be much weaker than

the initial demonstration suggested, (c) there may yet be important features that are not included in our paradigm that will elicit stronger malleability effects, and (d) these results have practical and conceptual implications for understanding malleability of implicit evaluations.

Malleability is real

Given the sheer volume of studies there is little doubt that Blair's (2002) conclusion about the malleability of implicit social cognition is correct. Even so, the published literature may overestimate the extent to which implicit social cognitions are malleable. Some other research also suggests that malleability of implicit cognition may be weaker than implied by the accumulated malleability literature. For example, Rydell and colleagues (Rydell & McConnell, 2006; Rydell, McConnell, Strain, Claypool, and Hugenberg, 2007) found that exposure to counterattitudinal examples had weak effects on implicit attitudes in an attitude formation paradigm.

Given that most psychological research is underpowered with small samples (Sedlmeier & Gigerenzer, 1989), researchers investigating malleability may be less likely to write-up results that fail to demonstrate malleability. If any one of our own studies, for example, had been done with a "typical" sample size, we might have dismissed the result and perhaps not pursued the other studies. With small samples, it is more likely that null effects are attributed to low power or quirks in the methodology rather than as indicating a lack of malleability. Our results cannot be dismissed so easily. With very large samples using an established paradigm, small or null effects are important indicators of the boundaries and limits of implicit malleability. It is possible that changes to our paradigm will elicit stronger effects. However, because we retained the "core features"

of the exposure manipulation; these results place important constraints on the interpretation and extent of implicit malleability.

Factors that may increase malleability

There were differences between the original paradigm and ours that we did not test systematically. For example, in the current experiments, participants viewed each exemplar twice (with the exception of Experiment 1) – once to identify their achievements, and once to identify their racial identity. DG duplicated both of these in their induction procedure. It is possible that the doubled exposure would increase the overall effect magnitude. However, we believe it unlikely that the effect size would increase by an order of magnitude to match the original demonstration.

Another factor that we did not examine is the amount of contact participants had with African Americans. In similar malleability research, Dasgupta and Rivera (2008) found that people with more contact to gay people were less sensitive to the malleability manipulation than people with less contact (and see Dasgupta and Asgari, 2004). This might indicate that malleability effects are stronger when the attitude is less elaborated. Except for Experiment 4 with University of Virginia students, there is not a particular reason to think that our samples had systematically more or less contact with African Americans than DG's. In any case, this could be an important factor for the magnitude of malleability effects.

Finally, as we have speculated, making the target category accessible may be a key component of the effect. DG, Dasgupta and Rivera (2008), and Dasgupta and Asgari (2004) all shared the joint features of exposure to admired exemplars and making the

target category highly accessible. Our last three experiments did include an accessibility manipulation, but strengthening it may further strengthen the malleability effect.

Are Malleability Effects Dependent on Demeaning the Dominant Group?

Brewer (1999) noted that liking one social group (e.g., White people) does not necessarily mean disliking another (e.g., Black people). Because the IAT is a relative measure of liking for Whites compared to Blacks, it is not clear whether Blacks are associated with negativity, or just comparatively less associated with positivity. Said another way, it is possible that both Blacks and Whites are already associated with the concept of "good", just with White individuals having a stronger association. Research with other implicit measures suggests that this possibility is viable (Nosek & Banaji, 2001; Bar-Anan, et al., 2009). If so, then exposure to positive Black individuals might not have much impact on shifting association strengths because the comparative distance between the positive exposure and the pre-existing association may be small. On the other hand, exposure to negative White individuals may be more impactful as the exemplars could be highly discrepant from the pre-existing positivity toward the group. From an applied perspective, this might be a discouraging possibility as it is rarely seen as good practice to *reduce* liking of one group in order to increase egalitarianism. Interestingly, Dasgupta and Asgari (2004) found that individuals were more likely to associate women with leadership after viewing admired female leaders *without* a negative comparison exposure. Likewise, Dasgupta and Rivera (2008) demonstrated that participants' anti-gay implicit bias can be reduced after viewing positive homosexual exemplars, without also showing negative hetereosexual exemplars. Nevertheless, it is

not clear whether and when exposure to positive group exemplars alone will be sufficient to change implicit cognitions about the target group.

Conclusion

That the effect of exposure to group members is a relatively weak influence on malleability of implicit cognitions may have important practical implications. Shifting implicit biases may not be as easy as implied by the existing experimental demonstrations. At the same time, our effect magnitude may be more "realistic" in the sense that it would be quite stunning if a 5-10 minute exposure to group members is sufficient to alter (temporarily or permanently) pre-existing associations accumulated over a lifetime. Real associative change may require persistent, even "permanent," exposure opportunities to group members that counter the social preferences that permeate the culture and mind.

References

Bar-Anan, Y., Nosek, B. A., & Vianello, M. (2009). The sorting paired features task: A measure of association strengths. *Experimental Psychology, 56,* 329-343.

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

Blair, I. V. (2002). The malleability of automatic stereotypes and prejudice. *Personality and Social Psychology Review, 6,* 242-261.

Brewer, M. B. (1999). The psychology of prejudice: Ingroup love or outgroup hate? *Journal of Social Issues, 55,* 429-444.

Dasgupta, N., & Asgari, S. (2004). Seeing is believing: Exposure to counterstereotypic women leaders and its effect on automatic gender stereotyping. *Journal of Experimental Social Psychology, 40,* 642-658.

Dasgupta, N., & Greenwald, A. G. (2001). On the malleability of automatic attitudes: Combating automatic prejudice with images of admired and disliked individuals. *Journal of Personality and Social Psychology, 81,* 800-814.

Dasgupta, N., & Rivera, L. M. (2008). When social context matters: The influence of long-term contact and short-term exposure to admired outgroup members on implicit attitudes and behavioral intentions. *Social Cognition, 26,* 112-123.

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

Evans, D. C., Garcia, D. G, Garcia, D. M., & Baron, R. S. (2003). In the privacy of their own homes: Using the internet to access racial bias. *Personality and Social Psychology Bulletin, 29,* 273-284.

Foroni, F., & Mayr, U. (2005). The power of a story: New, automatic associations from a single reading of a short scenario. *Psychonomic Bulletin & Review, 12,* 133-144.

Graham, J., Haidt, J., & Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology, 96,* 1029-1046.

Greenwald, T. G., McGhee, D. E., & Schwartz, J. K. 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.

Lindner, N. M., & Nosek, B. A. (2009). Alienable speech: Ideological variations in the application of free-speech principles. *Political Psychology, 30,* 67-92.

Lowery, B. S., Hardin, C. D., & Sinclair, S. (2001). Social influence effects on automatic racial prejudice. *Journal of Personality and Social Psychology, 81,* 842-855.

Mitchell, J. P., Nosek, B. A., & Banaji, M. R. (2003). Contextual variations in implicit evaluation. *Journal of Experimental Psychology: General, 132,* 455-469.

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

Nosek, B. A. (2005). Moderators of the relationship between implicit and explicit evaluation. *Journal of Experimental Psychology: General, 134,* 565-584.

Nosek, B. A., & Banaji, M. R. (2001). The go/no-go association task. *Social Cognition, 19,* 625-666.

Nosek, B. A., Greenwald, A. G., & Banaji, M. R. (2005). Understanding and using the Implicit Association Test: II. Method variables and construct validity. *Personality and Social Psychology Bulletin, 31(2),* 166-180.

Nosek, B. A., Greenwald, A. G., & Banaji, M. R. (2007). The Implicit Association Test at age 7: A methodological and conceptual review. In J. A. Bargh (Ed.), *Social Psychology and the Unconscious: The Automaticity of Higher Mental Processes* (pp. 265-292). New York: Psychology Press.

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

Nosek, B. A., Smyth, F. L., Hansen, J. J., Devos, T., Lindner, N. M., Ranganath, K. A., Smith, C. T., Olson, K. R., Chugh, D., Greenwald, A. G., & Banaji, M. R. (2007). Pervasiveness and correlates of implicit attitudes and stereotypes. *European Review of Social Psychology, 18,* 36-88.

Ranganath, K. A., & Nosek, B. A. (2008). Implicit attitude generalization occurs immediately; explicit attitude generalization takes time. *Psychological Science, 19,* 249-254.

Rudman, L. A., Ashmore, R. D. & Gary, M. L. (2001). "Unlearning" automatic biases: The malleability of implicit prejudice and stereotypes. *Journal of Personality and Social Psychology, 81,* 856-868.

Rydell, R. J., & McConnell, A. R. (2006). Understanding implicit and explicit attitude change: A systems of reasoning analysis. *Journal of Personality and Social Psychology, 91,* 995-1008.

Rydell, R. J., McConnell, A. R., Strain, L. M., Claypool, H. M., & Hugenberg, K. (2007). Implicit and explicit attitudes respond differently to increasing amounts of counterattitudinal information. *European Journal of Social Psychology, 37,* 867-878

Schneider, W., & Schiffrin, R. M. (1977). Controlled and automatic human information processing: I. Detection, search, and attention. *Psychological Review, 84,* 1-66.

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin, 105,* 309-316.

Sinclair, S., Lowery. B., Hardin, C. & Colangelo, A. (2005). Social tuning of automatic attitudes: The role of affiliative motivation. *Journal of Personality and Social Psychology, 89,* 583-592.

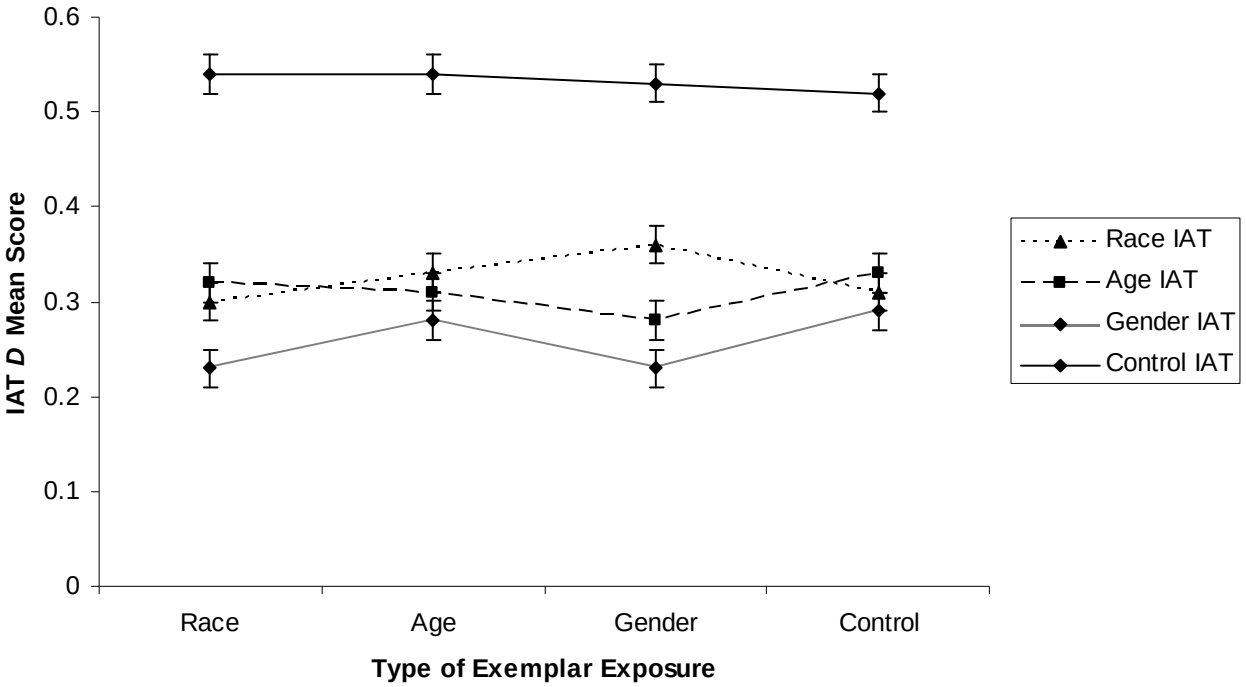
Skitka, L. J., & Sargis, E. G. (2005). Social psychological research and the Internet: The promise and peril of a new methodological frontier. In Y. Amichai-Hamburger (Ed), *In the social net: The social psychology of the Internet* (pp. 1-25). New York: Cambridge University Press.

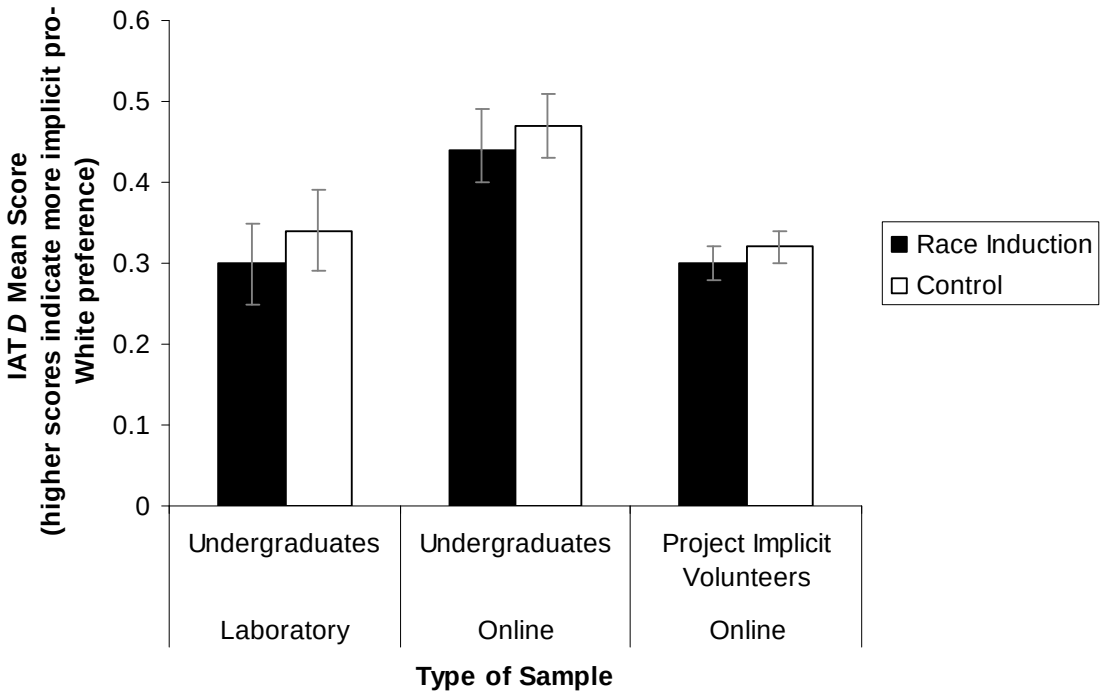
Wittenbrink, B., Judd, C. M., & Park, B. (2001). Spontaneous prejudice in context: Variability in automatically activated attitudes. *Journal of Personality and Social Psychology, 81*, 815-827.

Figure Captions

Figure 1. Implicit evaluations (IAT scores) by condition for Experiment 1.

Figure 2. Implicit evaluations (IAT scores) by condition for Experiment 3.


0.6
0.5
0.4
0.3
0.2
0.1
0
IAT D Mean Score
Race
Age
Gender
Control
Type of Exemplar Exposure
Race IAT
Age IAT
Gender IAT
Control IAT


IAT D Mean Score
(higher scores indicate more implicit pro-White preference)
0.6
0.5
0.4
0.3
0.2
0.1
0
Undergraduates
Laboratory
Undergraduates
Online
Project Implicit Volunteers
Online
Type of Sample
Race Induction
Control