Journal of Experimental Social Psychology 70 (2017) 8–18

Contents lists available at ScienceDirect

Journal of Experimental Social Psychology

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

Whites demonstrate anti-Black associations but do not reinforce them☆,☆☆

Jordan Axt *, Sophie Trawalter

University of Virginia, United States

HIGHLIGHTS

- Whites associate Black faces with negativity, but also value appearing unprejudiced.
- Whites completed a learning task pairing Blacks with positive or negative outcomes.
- Participants learned anti-Black associations quickly but did not strengthen them.
- Participants learned pro-Black associations more slowly but strengthened them.
- Results highlight how attitudes and prejudice concerns impact race-related behavior.

ARTICLE INFO

Article history:
Received 26 July 2016
Revised 17 November 2016
Accepted 25 November 2016
Available online xxxx

ABSTRACT

White people often associate Black people with negative information and outcomes. At the same time, many White people value not being or appearing prejudiced. In an inter-race context, these two forces may conflict. Whites may be better able to acquire anti-Black associations that align with their existing explicit or implicit attitudes, but may be unmotivated to strengthen these associations because they oppose their egalitarian values. Across five studies ($N > 1100$) including two pre-registered designs, Whites given a learning task were better able to initially acquire anti-Black racial associations but were unable or unwilling to then reinforce these associations. Conversely, Whites were less able to initially acquire pro-Black racial associations but then acquired and strengthened these associations. Finally, Whites were still unwilling or unable to reinforce anti-Black associations even when given a non-racial justification to do so. These results highlight the distinct but related influences of attitudes and prejudice concerns on race-related behavior.

White Americans, on average, associate Black people with negativity (Nosek, Banaji, & Greenwald, 2002), and have an easy time learning that Black people are associated with negative information. Not surprisingly then, they have an easier time associating negative experiences (e.g., an electric shock) with Black vs. White faces (Olsson, Ebert, Banaji, & Phelps, 2005). For example, in one study, participants received a mild electric shock while viewing either Black or White target faces. Subsequently, they viewed Black and White target faces without receiving the electric shocks. All the while, researchers measured participants' skin conductance responses as a proxy for fear, and found that White participants who had been shocked while viewing Black vs. White target faces then had an easier time learning to fear the Black vs. White target faces, and also had a harder time unlearning this association between Black faces and negative outcomes. This pattern of results is consistent with the notion that Whites have an easy time associating Blacks with negative information.

In the present work, we consider whether Whites also have an easier time associating Blacks with negativity when they are in control of seeking out information—information that could lead to anti-Black or pro-Black associations. There are reasons to believe that Whites may not have an easier time associating Blacks with negative information when they are in control of learning these associations. Biased information seeking and learning that Blacks are associated with negativity can signal that one is prejudiced, countering strong and widespread norms that racial bias is unacceptable. Many Whites value appearing and/or being unprejudiced (Plant & Devine, 1998), so much so that they avoid race-related discussions when possible (Apfelbaum, Norton, & Sommers, 2012; Trawalter & Richeson, 2008), and display their egalitarianism when threatened (Monin & Miller, 2001). For this reason,

☆ Article in press, *Journal of Experimental Social Psychology*.
☆☆ Author's note: Studies are numbered 1–5 for narrative style. Chronologically, studies were run in the following order: Study 1, Study 3a, Study 2, Study 3b, Study 4.
* Corresponding author at: Department of Psychology, University of Virginia, Box 400400, Charlottesville, VA 22904-4400, United States.
E-mail address: jaxt@virginia.edu (J. Axt).

http://dx.doi.org/10.1016/j.jesp.2016.11.013
0022-1031/

learning negative information about Black people becomes problematic; while it may be easier to do so since these anti-Black associations align with existing explicit and implicit attitudes, it may also be anxiety provoking from a self-perception or self-presentation perspective.

We explore the cognitive underpinnings of these two influences by investigating Whites' ability to acquire and reinforce information linking Black and White faces with negative and positive outcomes, or vice versa. Based on past work illustrating that people are better at learning racial associations that pair outgroup members with negative and ingroup members with positive information (e.g., Olsson et al., 2005), we might predict that Whites will acquire and more readily reinforce information that aligns with their pre-existing attitudes, specifically that Black is bad and White is good (e.g., Axt, Ebersole, & Nosek, 2014). Conversely, considering work that many White people are motivated to appear unprejudiced (e.g., Dunton & Fazio, 1997; Plant & Devine, 1998), we might predict that Whites will acquire and more readily reinforce information that presents themselves as unprejudiced towards Blacks, specifically that Black is good and White is bad.

To test this idea, we adapted the Iowa Gambling Task (IGT; Bechara, Damasio, Damasio, & Anderson, 1994), where participants select cards from four sets—two "good" sets where they earn points if chosen consistently and two "bad" sets where they lose points if chosen consistently. In our version, participants selected from good and bad sets but the sets had Black or White faces on them. The self-directed nature of task allows us to investigate what type of racial information is learned easiest (through more selecting of "good" sets) and also whether participants are able or willing to then seek out the racial information that will improve performance. Notably, we incentivized performance, meaning participants sacrificed possible rewards by not acquiring or reinforcing certain racial associations.

The self-guided nature of the task departs from previous training interventions that sought to either reduce (or strengthen) automatic racial bias. For instance, prior work has illustrated that actively negating stereotype-consistent information in a training session decreased automatic racial bias (Johnson, Kopp, & Petty, 2016; Kawakami, Dovidio, Moll, Hermsen, & Russin, 2000), and negating stereotype-inconsistent information increased automatic racial bias (Johnson et al., 2016). Likewise, being exposed to counter-stereotypic exemplars or undergoing training that paired Black faces with positive and White faces with negative words reduced automatic racial bias, at least when tested immediately (Lai et al., 2016). In these cases, the training manipulations were uniformly imposed upon participants with no room for participants to actively decide how much stereotype-consistent or stereotype-inconsistent information they wanted to seek out. However, in our studies, we merely designed a learning context that either did or did not align with pre-existing racial attitudes and investigated the extent to which participants displayed associations between race and positive or negative outcomes.

Across five studies, the IGT paradigm revealed an intriguing asymmetry in participants' acquiring and strengthening of racial associations. White participants readily demonstrated an association that paired Black faces with negative outcomes, supporting earlier work showing that Whites are better at learning an association between racial outgroups and aversive stimuli (Olsson et al., 2005). However, Whites in our studies were then unwilling or unable to strengthen this initial anti-Black association. Conversely, White participants were less able to initially acquire an association that paired Black faces with positive and White faces with negative outcomes, but were willing and able to reinforce this association. By using a more self-guided paradigm, this work sheds light on the degree of control that individuals may have over the associations they initially learn and later reinforce by revealing the dual influence of pre-existing attitudes as well as motivations to appear unprejudiced on race-related behaviors.

In Study 1, we tested this idea of whether participants differed in their ability to initially learn and then strengthen anti-Black versus anti-White associations.

1. Study 1

1.1. Methods

1.1.1. Participants

For all studies, we report how we determined our sample size and report all manipulations and measures. In Study 1, we sought to collect 50 White, American, native English speakers for each experimental condition. Due to group data collection and random assignment to conditions, the final sample consisted of 176 (57.4% female, $M_{Age} = 19.1$) undergraduates who participated in exchange for partial course credit.

1.1.2. Procedure

Participants completed the study at individual computer carrels with 0 to 3 other participants in the room. After providing consent, participants completed measures in the following order: modified Iowa Gambling Task, measures of explicit racial attitudes and motivations in a randomized order, a measure of implicit racial attitudes, and a demographics survey. Participants were then debriefed and given feedback on their implicit task performance (see https://osf.io/b762g/ for online supplements, materials, and data from all studies).

1.1.2.1. Iowa Gambling Task. Participants completed a modified version of the IGT. In an IGT, participants complete trials where they select from one of four card sets on the computer screen. After selecting, participants either win or lose points. The sets differ in regards to whether they produce gains or losses in points on average. Over the course of the task, it is the participant's job to earn as many points as possible.

In this version of the IGT, there were two "winning" and two "losing" sets. In the winning sets, participants gained an average of 25 points per trial. In the two losing sets, participants lost an average of 25 points per trial. In both the winning and losing sets, 60% of trials ended with point gains and 40% with point losses, but the losing sets had larger losses than the winning sets. More information about task scoring can be found in the online supplement.

The task lasted for 120 trials. Participants started with 2000 points. Participants' point total was displayed on screen throughout the task, and they received feedback regarding how many points they won or lost after every trial. Participants were assigned to one of three experimental conditions. In the Card Decks condition ($n = 51$), all sets were of an image of the back of a playing card. In the two race conditions, the card deck images were replaced by images of Black and White people. In the Black = Bad condition ($n = 67$), winning sets were comprised of White faces and losing sets comprised of Black faces (for each race, images were selected without replacement from a pool of 30 faces and stimuli were not repeated between sets). In the Black = Good condition ($n = 58$), winning sets were comprised of Black faces and losing sets comprised of White faces.[1]

To incentivize good performance, participants in the top 10% of points earned within the study received a $10 gift card. Specifically, participants were instructed:

> In this experiment, you will be asked to repeatedly select a [face/card] from one of four sets. You can select a [face/card] from the set by clicking the mouse on the [face/card] you want to select.
>
> With each [face/card], you can win some points, but you can also lose some. Some sets will be more profitable than others. Try to choose [faces/cards] from the most profitable sets so that your total winnings will be as high as possible. All participants will receive participation credit, but those participants in the highest 10% of points earned will be given a $10 gift card.

[1] We chose this naming instead of *White = Good* vs. *White = Bad* or *Attitude Consistent* vs. *Attitude Inconsistent* because Black faces were likely the most salient aspect of the task.

You will get 120 chances to select a [face/card] from the set that you think will give you the highest winnings.

1.1.2.2. Explicit racial attitudes and motivations. After the IGT, participants completed measures of racial attitudes and motivations to investigate how each influences performance on the task. For explicit racial attitudes, participants completed a single-item measure of explicit preference for White relative to Black people (−3 = I strongly prefer Black people to White people, +3 = I strongly prefer White people to Black people). For race-related motivations, we participants completed the 10-item measure of internal motivation (IMS) and external motivation (EMS) to control prejudice (Plant & Devine, 1998).

1.1.2.3. Implicit racial attitudes. To measure implicit racial attitudes, participants completed a seven-block Implicit Association Test (IAT; Greenwald, McGhee, & Schwartz, 1998) measuring the strength of the association between the concepts "Good" and "Bad" and the categories "White American" and "Black American". IAT responses were scored by the *D* algorithm (Greenwald, Nosek, & Banaji, 2003), such that more positive scores reflected a stronger association between White American and good and Black American and bad. The procedure followed the recommended design and exclusion criteria from Nosek, Greenwald, and Banaji (2007). See the Appendix for more information on the IAT procedure.

1.1.2.4. Demographics. Participants completed a nine-item demographics questionnaire reporting information such as gender, age, race, country of birth and native language.

1.2. Results

IAT *D* scores ($M = 0.36$, $SD = 0.39$, Cohen's *d* versus 0 = 0.92) and the explicit attitude item ($M = 0.57$, $SD = 0.68$, Cohen's *d* versus 0 = 0.84) indicated more positive evaluations for White relative to Black people. Participants also displayed higher levels of internal ($M = 7.25$, $SD = 1.15$) than external ($M = 5.64$, $SD = 1.54$) motivation to respond without prejudice. These variables did not reliably differ across experimental conditions (all F's < 1.55, all p's > 0.216).

1.2.1. IGT performance

We analyzed performance on the IGT through hierarchical linear model (HLM), as trials were nested within participants. We used HLM because it offered the most fine-grained and powerful approach, though there are other ways to analyze these data. For instance, another option could have been to divide the IGT into four 30-trial blocks and test for improvement within conditions over time across blocks, as well as testing for differences between conditions at each block. Indeed, we ran these analyses for all studies and they reach the same conclusions as the HLM analyses reported in the main text. Block analyses are available in the online supplement.

For all studies, we first recoded trial number to be on a 0–119 scale (to allow for an interpretable intercept) and then divided trial number by 119 so the first trial had a value of 0 and the last trial a value of 1 (to allow for more interpretable regression coefficients). Both slopes and intercepts were modeled as random factors. Because the outcome was binary, we used a Bernoulli model. We also used robust maximum likelihood estimation to adhere to best practice guidelines (Kline, 2011).

We tested whether conditions differed both in terms of their *intercept* (i.e., the predicted likelihood of selecting from a winning set on the first trial) and their *slope* (i.e., changes in the predicted likelihood of selecting from a winning set over the course of the task). An unreliable intercept coefficient (not reliably different from zero) would indicate a model predicting that participants' initial selection was no different from chance. A positive intercept coefficient would indicate a model predicting that participants were more likely than chance to select from a winning set initially, and a negative intercept coefficient would indicate a model predicting that participants were less likely than chance to select from a winning set initially. An unreliable (not reliably different from zero) coefficient for slope would indicate a model predicting that participants did not change in the likelihood of selecting from winning sets over the course of the task. A positive or negative coefficient for the slope would indicate a model predicting an increase or decrease, respectively, in the likelihood of selecting from winning sets over the course of the task.

We interpreted the intercept as representing initial associations and the slope as representing how well associations were acquired or strengthened throughout the task. Importantly, these analyses present model estimates; that is, predictions on how best to understand task performance in general. The intercept is not necessarily the average of participants' actual choices on the first trial but the predicted choice on the first trial given performance throughout the task. For instance, even if the first few trials showed near-chance performance that quickly gave rise to above-chance selections that then did not change, the HLM analyses would show a positive intercept, since there was not enough meaningful variation in task performance over time to merit a reliable slope. Conversely, if initial performance was at chance levels but only gradually increased to above-chance levels throughout the task, the HLM analyses would show an unreliable (no different from zero) intercept and positive slope, as there was enough meaningful variation in performance over time to merit differing estimates for the first trial and for performance throughout the task. As a result, we used intercept as an indicator of early associations and slope as an indicator of how associations did or did not strengthen throughout the task (see Greenwald, Klinger, & Schuh, 1995 for a similar treatment of dissociating intercept and slope to reveal separate psychological processes).

From these two estimates, there are then different inferences that can be made from the four combinations of a positive vs. unreliable slope vs. intercept. An unreliable intercept and unreliable slope would suggest that participants failed to acquire an association between which sets were good and which were bad, both initially and as the task progressed. A positive intercept and positive slope would suggest that participants formed an initial association between which sets were good and which were bad, and that this association was then strengthened as the task progressed. An unreliable intercept but a positive slope would suggest that participants failed to learn the initial association between which sets were good and which were bad, but that this association was acquired as the task progressed.

Finally, a positive intercept but unreliable slope would suggest that participants formed the initial association between which sets were good and which were bad, but that this association was not further strengthened as the task progressed. A positive intercept but unreliable slope would suggest that participants were able to learn an association between certain sets and positive or negative outcomes initially in the task, but the strength of this association did not reliably change across trials and as the task progressed. That is, a positive intercept and unreliable slope indicates a model where the most appropriate understanding of participant responses illustrated above-chance performance (i.e., understanding associations between sets and task outcomes) that did not meaningfully change throughout the task (no reliable slope).

We first analyzed the Card Decks condition, predicting the likelihood of selecting from a winning set by the Level 1 variables of an intercept and trial number with no additional Level 2 variables. The intercept was not reliable ($B = 0.05$, $t = 1.01$, $p = 0.320$, OR = 1.06, 95% CI [0.95, 1.18]), but the slope was positive ($B = 0.32$, $t = 2.36$, $p = 0.022$, OR = 1.37, 95% CI [1.05, 1.80]). Here, the intercept indicates that on the first trial, participants were equally likely to initially pick from winning or losing sets, which makes sense given that participants in the Card Decks condition had no information on the first trial to differentiate between good and bad sets. However, the positive slope means participants were more likely to select from good sets as the task progressed, indicating they formed an association between certain sets and point gains or losses.

Next, we analyzed performance within the two race conditions. We predicted the likelihood of selecting from a winning set by the Level 1 variables of an intercept and trial number and the Level 2 variable of experimental condition (0 = Black = Bad, 1 = Black = Good). Relative to the Black = Bad condition, participants in the Black = Good condition had a lower intercept ($B = -0.45$, $t = 3.73$, $p < 0.001$, OR = 0.64, 95% CI [0.50, 0.81]) but a more positive slope ($B = 0.63$, $t = 3.47$, $p = 0.001$, OR = 1.88, 95% [1.31, 2.69]).

A model looking only at the Black = Bad condition revealed a positive intercept ($B = 0.30$, $t = 3.57$, $p = 0.001$, OR = 1.35, 95% CI [1.14, 1.59]) but no reliable slope ($B = 0.04$, $t = 0.39$, $p = 0.699$, OR = 1.04, 95% CI [0.85, 1.27]). Conversely, a final model looking only at the Black = Good condition revealed a negative intercept ($B = -0.17$, $t = 2.02$, $p = 0.047$, OR = 1.04, 95% CI [0.71, 0.998]) but a positive slope ($B = 0.76$, $t = 4.39$, $p < 0.001$, OR = 2.14, 95% CI [1.51, 3.03]). See Fig. 1 for a graphical display of all conditions.

As Black = Good was the only race condition to show a positive slope, we then tested whether the intercept and slope were moderated by IMS, EMS, IAT *D* score, and the explicit preference item (all standardized). Full moderation analyses for all conditions are available in the online supplement. Within each study, we report only the statistically significant moderators for that sample, but later meta-analyze results from all studies to provide the most precise estimates for each moderator.

Within the Black = Good condition in Study 1, explicit racial preference was associated with a lower intercept ($B = -0.18$, $t = 2.16$, $p = 0.035$, OR = 0.83, 95% CI [0.70, 0.99]) and steeper slope ($B = 0.42$, $t = 2.43$, $p = 0.019$, OR = 1.52, 95% CI [1.08, 2.14]), indicating that

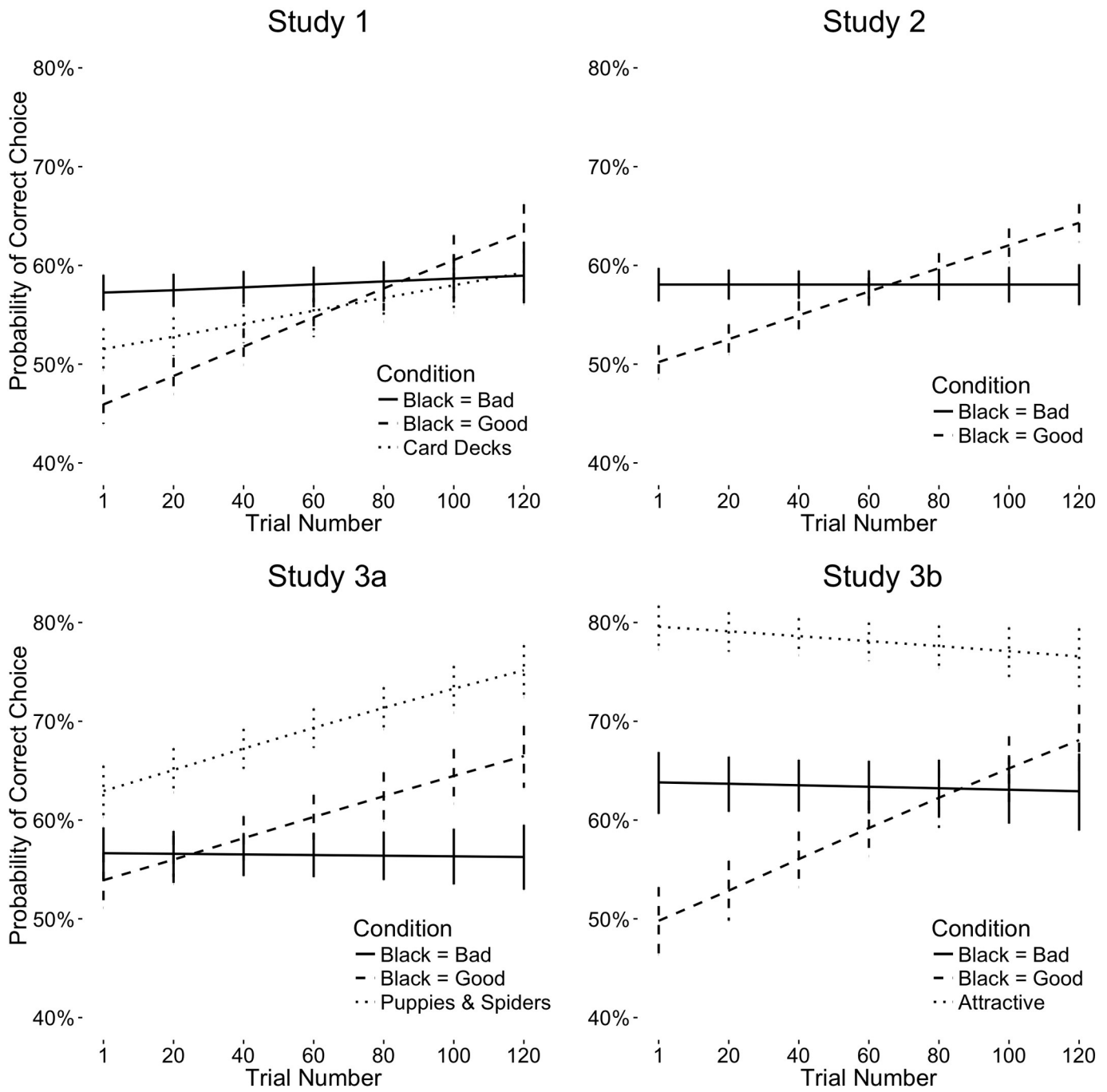

Fig 1. Estimated probability of making a correct choice (selecting from a winning set) throughout the task for each condition within each study. Error bars represent standard errors at various intervals of the task.

participants higher in explicit preference for Whites over Blacks were less likely to select Black faces from winning sets initially, but then showed greater changes in the likelihood of selecting (correct) Black faces over the course of the task.

1.3. Discussion

In the Card Decks condition, participants did not distinguish good from bad sets initially, but learned to associate good sets with positive and bad sets with negative outcomes during the task. When including racial information, participants initially associated Black faces with negative outcomes in the Black = Bad condition (a positive intercept), but were unable or unwilling to strengthen this association (an unreliable slope). Conversely, participants did not initially learn an association between Black faces and positive outcomes in the Black = Good condition (a negative intercept), but acquired this association during the task (a positive slope).

The positive and negative intercepts in the race conditions may seem striking. How could our models predict that participants would perform above or below chance on the first trial? We believe the answer is partly due to the greater likelihood of White faces being selected first, regardless of condition, a bias that may have arisen due to pre-existing attitudes favoring Whites over Blacks. Across all studies, White faces were preferred on the first trial in the Black = Bad (56.3%) and Black = Good (56.2%) conditions. That is, the Black = Bad conditions and Black = Good conditions did not differ in the likelihood of selecting a White face on the first trial.

However, selecting White faces was a winning strategy only in the Black = Bad condition. There, selecting White faces was reinforced, and participants then continued selecting White faces, allowing them to quickly learn to associate White with positive and Black with negative outcomes. Moreover, the flat *slope* in the Black = Bad condition suggests that this initial anti-Black association remained present throughout the task. If participants had merely selected White faces initially but never formed an association between race and task outcomes, we would expect a negative slope as performance returned to chance levels; instead, the positive intercept and flat slope suggests that participants learned to pair Black faces with negative outcomes quickly, but the strength of this association did not change as the task progressed. In other words, the HLM analyses suggested that the simplest prediction of performance in the Black = Bad conditions was of learning associations between race and task outcomes in a way that was present throughout the entire task. Conversely, selecting White faces initially was not a winning strategy in the Black = Good condition, but the positive slope indicates that participants increased their selection of good (Black) faces across the task, suggesting that they were able and willing to acquire and strengthen this pro-Black association during the task.

In Study 2, we attempted to manipulate prejudice concerns. If prejudice concerns influenced the ability and/or willingness to acquire or strengthen racial associations, then increasing prejudice concerns should make people less willing to reinforce associations in the Black = Bad condition and more willing in the Black = Good condition. In Study 2, we sought to both replicate Study 1 and examine whether heightening prejudice concerns would impact the ability to acquire and strengthen anti-Black or anti-White associations.

2. Study 2

2.1. Methods

2.1.1. Participants

We sought to collect at least 60 White, American, native English speakers for each experimental condition. Due to group data collection and random assignment to conditions, the final sample consisted of 267 (74.5% female, $M_{Age} = 18.5$) undergraduates who participated in exchange for partial course credit.

2.1.2. Procedure

The experimental design was the same as Study 1 except for the following changes. First, to investigate whether differences in IGT performance across conditions could be explained by participants finding the task more difficult or through exerting more effort, participants completed two items measuring how difficult the task was (1 = Not at all, 5 = Extremely) and how much effort they gave during the task (1 = Did not try at all, 5 = Tried extremely hard) immediately after completing the IGT. Second, Study 2 had no control condition; participants were randomly assigned to either the Black = Good or Black = Bad version of the IGT. To heighten concerns about appearing prejudiced, participants were randomly assigned to complete the IAT before (vs. after) the IGT, a manipulation that has been used in previous research to heighten prejudice concerns (e.g., Effron, Miller, & Monin, 2012; Merritt et al., 2012). To strengthen the threat manipulation, participants also read text about the logic behind the IAT, which highlighted how the test can often reveal automatic racial biases that are not consciously endorsed but may still influence behavior. Only participants completing the IAT before the IGT read the extra information about how the IAT works as a measure of racial bias.

2.2. Results

2.2.1. Manipulation check

We submitted IMS and EMS responses to a 2 (Task version: Black = Good vs. Black = Bad) by 2 (Order: IAT before vs. IAT after) ANOVA. These analyses revealed no reliable main effects of task version, IAT order, or interaction between the two variables, all F's < 2.19, all p's > 0.140. For this reason, we considered our attempt at manipulating motivations to appear prejudiced unsuccessful,[2] and the analyses reported below collapse across the threat manipulation (Black = Bad $n = 130$; Black = Good $n = 137$). Analyses that include the threat manipulation are available in the online supplement.

2.2.1.1. Attitude and motivation measures. Participants reported finding the task to be relatively difficult ($M = 3.77$, $SD = 0.85$) and trying moderately hard ($M = 2.70$, $SD = 0.91$). IAT D scores ($M = 0.44$, $SD = 0.41$; $d = 1.07$) and explicit attitudes ($M = 0.64$, $SD = 0.77$ $d = 0.83$) indicated more positive evaluations of White relative to Black people. Participants again displayed higher levels of internal ($M = 7.48$, $SD = 1.19$) than external ($M = 5.72$, $SD = 1.46$) motivation to respond without prejudice. These variables did not reliably differ across experimental conditions (all t's < 1.44, all p's > 0.152).

2.2.1.2. IGT performance. Task performance was analyzed using the same method as Study 1. Replicating Study 1, relative to the Black = Bad condition, participants in the Black = Good condition had a lower intercept ($B = -0.32$, $t = 3.30$, $p = 0.001$, OR = 0.73, 95% CI [0.60, 0.88]) but a more positive slope ($B = 0.57$, $t = 4.16$, $p < 0.001$, OR = 1.77, 95% CI [1.35, 2.32]).

A model looking only at the Black = Bad condition revealed a positive intercept ($B = 0.33$, $t = 4.91$, $p < 0.001$, OR = 1.38, 95% CI [1.22, 1.58]) but not a reliable slope ($B = -0.02$, $t = -0.19$, $p = 0.852$, OR = 0.98, 95% CI [0.83, 1.17]). Finally, a model looking only at the Black = Good condition revealed no reliable intercept ($B = 0.005$, $t = 0.07$, $p = 0.943$, OR = 1.005, 95% CI [0.88, 1.15]) but a positive slope ($B = 0.59$, $t = 5.58$, $p < 0.001$, OR = 1.80, 95% CI [1.46, 2.21]). See Fig. 1 for a graphical display of all conditions.

As the Black = Good condition was the only race-based condition to show a positive slope, we tested whether the intercept and slope were moderated by effort, perceived difficulty, IMS, EMS, IAT D score, and the explicit preference item (all standardized). Full analyses for all conditions are available in the online supplement. As in Study 1, only

[2] This manipulation may have failed to alter IMS or EMS scores because the racial IGT itself raised prejudice concerns.

explicit preference was a reliable moderator of intercept ($B = -0.23$, $t = 2.83$, $p = 0.006$, OR = 0.80, 95% CI [0.68, 0.93]), indicating that participants higher in explicit preference for Whites over Blacks were less likely to select a (correct) Black face initially. None of these variables were reliable moderators of slope (all t's < 1.49, all p's > 0.139).

2.3. Discussion

While the threat manipulation did not change prejudice motivations, results replicated Study 1. Participants in the Black = Bad condition learned to associate Black faces with negative outcomes initially, but this association was not strengthened. Participants in the Black = Good condition did not associate Black faces with positive outcomes initially, but acquired this association as the task progressed.

One alternative interpretation for Studies 1–2 is a ceiling effect for the Black = Bad condition. In both race conditions, participants ended with comparable probabilities of selecting from a winning set on the final trial. As a result, it is possible that since participants in the Black = Bad condition started with a greater likelihood of selecting winning sets, they could not strengthen this racial association because initial performance was the best allowed given the task's difficult nature.

To address this possibility, participants in Study 3a completed a version of the task where we believed they would be able and willing to both learn and strengthen an association, using puppies and spiders instead of Black and White faces. In addition, to test the specificity of this inability to strengthen the association in the Black = Bad condition, Study 3b tested whether participants would perform similarly when the task involved another social dimension: physical attractiveness. In Study 3a and 3b, we then investigated how well participants could acquire and strengthen both racial and non-racial associations.

3. Study 3a & 3b

3.1. Methods

3.1.1. Participants

In Study 3a, we sought to collect at least 70 White, American, native English speakers for each condition. Due to group data collection and random assignment to conditions, the final sample consisted of 229 (47.2% female, $M_{Age} = 19.1$) undergraduates who participated in exchange for partial course credit (Puppies and Spiders $n = 77$; Black = Bad $n = 80$; Black = Good $n = 72$).

In Study 3b, we sought to collect at least 60 eligible participants for each condition. The final sample consisted of 187 White, American, native English speakers (58.8% female, $M_{Age} = 18.6$) who participated in exchange for partial course credit (Attractive $n = 66$; Black = Bad $n = 61$; Black = Good $n = 60$). The pre-registration and analysis plan for Study 3b can be found at https://osf.io/q2vwd/.

3.1.2. Procedure

The experimental design was the same as Study 2 except for the following changes. First, we removed the threat manipulation. Second, we added an extra condition to each study. In Study 3a, participants in the Puppies and Spiders condition completed a version of the task where winning sets consisted of images of puppies and losing sets consisted of images of spiders (for each animal, stimuli were selected without replacement from a pool of 30 images and stimuli were not repeated between sets).

In Study 3b, we replaced the Puppies and Spider condition with the Attractiveness condition, where winning sets were paired with faces that had been previously pretested (Axt et al., under review) as being higher in physical attractiveness and losing sets were paired with faces rated as lower in physical attractiveness. Faces were all White and consisted of both males and females (for each level of attractiveness, stimuli were selected without replacement from a pool of 30 images split evenly between genders, and stimuli were not repeated between sets). Within the Attractiveness condition, explicit and implicit attitude measures were altered to preferences for more and less physically attractive people, and the IMS and EMS scales were adapted to be about prejudice towards more and less physically attractive people.

3.2. Results

3.2.1. Attitude and motivation measures

In Study 3a, participants reported finding the task to be relatively difficult ($M = 3.79$, $SD = 0.83$) and trying moderately hard ($M = 2.79$, $SD = 0.99$). IAT D scores ($M = 0.37$, $SD = 0.38$, $d = 0.97$) and the explicit attitude item ($M = 0.51$, $SD = 0.69$, $d = 0.74$) indicated more positive evaluations for White relative to Black people. Participants again displayed higher levels of internal ($M = 7.32$, $SD = 1.22$) than external ($M = 5.55$, $SD = 1.60$) motivation to respond without prejudice. These variables did not reliably differ across experimental conditions (all F's < 1.55, all p's > 0.216), except for the IAT, $F(2226) = 3.21$, $p = 0.042$. Participants in the Black = Good condition ($M = 0.29$, $SD = 0.37$, $d = 0.78$) had lower IAT D scores than participants in the Puppies and Spiders condition ($M = 0.44$, $SD = 0.37$, $d = 1.19$), $t(147) = 2.56$, $p = 0.011$, $d = 0.42$, 95% CI [0.09, 0.74], though there were no reliable differences in IAT D scores between the two race conditions, $t(150) = 1.20$, $p = 0.232$, $d = 0.20$, 95% CI [−0.12, 0.51].

In Study 3b, participants also reported finding the task to be relatively difficult ($M = 3.63$, $SD = 0.83$) and trying moderately hard ($M = 2.88$, $SD = 0.85$). There were no reliable differences in reported task difficulty across conditions, $F(2184) = 0.36$, $p = 0.698$, but there were on reported effort, $F(2184) = 3.65$, $p = 0.028$. Participants in the Black = Bad condition ($M = 2.67$, $SD = 0.85$) reported lower effort than those in the Attractive condition ($M = 3.08$, $SD = 0.77$), $t(125) = 2.81$, $p = 0.006$, $d = 0.50$, 95% CI [0.14, 0.85], though there were no reliable differences in effort between the two race conditions, $t(119) = 1.32$, $p = 0.188$, $d = 0.24$, 95% CI [−0.12, 0.60].

For Study 3b, participants in the two race conditions had IAT D scores ($M = 0.42$, $SD = 0.36$, $d = 1.17$) and explicit attitudes ($M = 0.55$, $SD = 0.82$, $d = 0.67$) that indicated both an explicit and implicit preference for Whites over Blacks. Participants again displayed higher levels of internal ($M = 7.29$, $SD = 1.18$) than external ($M = 5.77$, $SD = 1.49$) motivation to respond without prejudice. These variables did not reliably differ between conditions (all t's < 0.94, all p's > 0.349), except for a small effect in the IAT, where participants in the Black = Good condition had lower IAT scores ($M = 0.36$, $SD = 0.37$) than participants in the Black = Bad condition ($M = 0.49$, $SD = 0.36$), $t(119) = 2.10$, $p = 0.038$, $d = 0.38$, 95% CI [0.02, 0.74].

Participants in the Attractiveness condition in Study 3b had IAT D scores ($M = 0.85$, $SD = 0.27$, $d = 3.15$) and explicit attitudes ($M = 1.41$, $SD = 0.82$, $d = 1.72$) that indicated preferences for more physically attractive people. Participants also displayed higher levels of internal ($M = 6.72$, $SD = 1.23$) than external ($M = 6.08$, $SD = 1.23$) motivation to respond without prejudice.

3.2.2. IGT performance

We analyzed the IGT in the same manner as previous studies. In the Puppies and Spiders condition in Study 3a, both the intercept ($B = 0.55$, $t = 4.15$, $p < 0.001$, OR = 1.73, 95% CI [1.33, 2.44]) and slope ($B = 0.61$, $t = 2.80$, $p = 0.007$, OR = 1.85, 95% CI [1.19, 2.86]) were positive. In the Attractiveness condition in Study 3b, there was a highly positive intercept ($B = 1.41$, $t = 7.93$, $p < 0.001$, OR = 4.10, 95% CI [2.88, 5.86]) but not slope ($B = -0.29$, $t = 1.64$, $p = 0.106$, OR = 0.75, 95% CI [0.52, 1.07]).

Next, we analyzed performance within the two race conditions. For Study 3a, relative to the Black = Bad condition, participants in the Black = Good condition did not differ in intercept ($B = -0.12$, $t = 0.89$, $p = 0.372$, OR = 0.88, 95% CI [0.68, 1.16]) but had a more positive slope ($B = 0.57$, $t = 2.89$, $p = 0.005$, OR = 1.77, 95% CI [1.20, 2.62]). In

Study 3b, relative to the Black = Bad condition, participants in the Black = Good condition had a lower intercept ($B = -0.56$, $t = 3.64$, $p = 0.001$, OR = 0.57, 95% CI [0.42, 0.77]) but a more positive slope ($B = 0.78$, $t = 2.69$, $p = 0.009$, OR = 2.18, 95% CI [1.23, 3.87]).

A model looking only at the Black = Bad conditions revealed a positive intercept (Study 3a: $B = 0.26$, $t = 2.61$, $p = 0.011$, OR = 1.30, 95% CI [1.06, 1.59]; Study 3b: $B = 0.55$, $t = 5.19$, $p < 0.001$, OR = 1.73, 95% CI [1.40, 2.14]) but not a reliable slope (Study 3a: $B = -0.05$, $t = -0.42$, $p = 0.677$, OR = 0.95, 95% CI [0.73, 1.22]; Study 3b: $B = -0.04$, $t = -0.22$, $p = 0.830$, OR = 0.96, 95% CI [0.66, 1.40]). Finally, a model looking only at the Black = Good condition revealed no reliable intercept (Study 3a: $B = 0.12$, $t = 1.34$, $p = 0.185$, OR = 1.13, 95% CI [0.94, 1.36]; Study 3b: $B = -0.02$, $t = -0.15$, $p = 0.882$, OR = 0.98, 95% CI [0.79, 1.23]) but a positive slope (Study 3a: $B = 0.57$, $t = 3.64$, $p = 0.001$, OR = 1.77, 95% CI [1.30, 2.41]; Study 3b: $B = 0.78$, $t = 3.51$, $p = 0.001$, OR = 2.19, 95% CI [1.40, 3.43]). See Fig. 1 for a graphical display of all conditions.

As the Black = Good condition was the only race condition to show a positive slope in either study, we tested whether the intercept and slope were moderated by effort, perceived difficulty, IMS, EMS, IAT *D* score, and the explicit preference item (all standardized). Full analyses for all conditions are available in the online supplement. None of these variables were reliable moderators of intercept in Study 3a (all t's < 1.67, all p's > 0.100), but explicit preference was a reliable moderator of intercept in Study 3b ($B = -0.31$, $t = 2.59$, $p = 0.013$, OR = 0.74, 95% CI [0.58, 0.93]), indicating that participants higher in explicit preference for Whites over Blacks were less likely to select Black faces from winning sets early in the task. Only explicit preference was a reliable moderator of slope in Study 3a ($B = 0.40$, $t = 2.41$, $p = 0.019$, OR = 1.49, 95% CI [1.07, 2.06]), indicating that participants higher in explicit preference for Whites over Blacks showed greater changes in the likelihood of selecting Black faces over the course of the task, but none of these variables were reliable moderators of slope in Study 3b (all t's < 1.52, all p's > 0.134).

3.3. Discussion

Replicating Studies 1 and 2, participants in the Black = Bad conditions acquired the association between Black faces and negative outcomes initially, but this association was not strengthened. Participants in the Black = Good conditions did not learn the association between Black faces and positive outcomes initially, but acquired this association as the task progressed. Participants in the Puppies and Spiders condition were able to associate puppies with positive outcomes and spiders with negative outcomes initially, *and* strengthen this association throughout the task. These results, a positive intercept and slope in the Puppies and Spiders condition, suggest that the lack of positive slope in the Black = Bad conditions is not a ceiling effect but rather an unwillingness or inability to strengthen the initial association between Black faces and negative outcomes, as participants could both learn an association initially and strengthen it when completing an IGT with different stimuli.

Participants in the Attractiveness condition showed a strong association between attractive faces and positive outcomes initially, but did not strengthen this association. The lack of strengthening in the Attractiveness condition may be due to the high rate of selecting attractive faces initially (80.3% of participants selected an attractive face first). This performance level may represent an actual ceiling effect, as participants in the Puppies and Spiders condition arrived at a comparable rate of correct responses by the end of the task. However, it's possible that the flat slope in the Attractiveness condition reveals a more general inability or unwillingness to reinforce associations that may carry self-presentation concerns.

Study 4 tests the robustness of the inability or unwillingness to associate Black faces with negative outcomes by seeing if similar performance emerges after providing a non-racial justification to develop negative associations towards Blacks. Participants completed a version of the Black = Bad IGT but were told that the White faces belonged to moral and the Black faces to immoral people. This manipulation would allow us to investigate whether participants would be willing or able to associate Black faces with negative outcomes after being given a clear and unambiguous reason to do so. In Study 4, we then investigated whether participants would be willing to strengthen anti-Black associations if given an external rationalization for doing so.

4. Study 4

4.1. Methods

4.1.1. Participants

We sought to collect at least 80 White, American, native English speakers for each of the three experimental conditions. Due to random assignment to conditions, the final sample consisted of 253 (69.6% female, $M_{Age} = 19.2$) undergraduates who participated in exchange for partial course credit. The study's pre-registration and analysis plan can be found at https://osf.io/rwt7a/.

4.1.2. Procedure

4.1.2.1. IGT. Participants were randomly assigned to complete one of three versions of the racial IGT. The Black = Bad condition ($n = 80$) had the same design as that used in previous studies. There were two additional conditions. These two conditions received more information about the faces being used in the task. During the instructions for the task, participants were informed that the faces they were going to see came from seminary school students. The seminary had both "good" students, who were excellent members of their community and would be graduating soon, and "bad" students, who frequently engaged in dishonest behavior and were in the process of being expelled (see the online supplement for the full text). During the task, each set was labeled as containing either "good" or "bad" students, and participants were again told to try to select faces from sets that would earn them the most points. In these experimental conditions, the two winning sets always consisted of "good" students and the two losing sets always consisted of "bad" students.

In the Black = Bad + Context condition ($n = 90$), the two winning sets consisted of White faces, and the two losing sets consisted of Black faces. In the All Black + Context condition ($n = 83$), all four sets consisted of Black males using the same images as previous studies.

4.1.2.2. Other measures. After the learning task, participants completed the previously used measures in the following order: perceptions of task difficulty and effort, explicit racial preferences, IMS and EMS, a White-Black evaluative IAT, and the nine-item demographics questionnaire.

4.2. Results

Participants found the task to be difficult ($M = 3.78$, $SD = 0.84$) and tried moderately hard ($M = 2.89$, $SD = 0.97$). There were reliable differences across conditions on reported task difficulty $F(2, 250) = 4.99$, $p = 0.001$ and reported effort, $F(2250) = 3.24$, $p = 0.041$. Participants in the All Black + Context condition reported higher effort ($M = 4.02$, $SD = 0.75$) and greater difficulty ($M = 2.96$, $SD = 1.03$) than participants in the Black Bad + Context (Difficulty: $M = 3.54$, $SD = 0.91$; Effort: $M = 2.60$, $SD = 0.93$), condition (Difficulty: $t(171) = 3.76$, $p < 0.001$, $d = 0.57$, 95% CI [0.27, 0.88]; Effort: $t(171) = 2.44$, $p = 0.016$, $d = 0.37$, 95% CI [0.07, 0.67]).

IAT *D* scores ($M = 0.40$, $SD = 0.37$, $d = 1.08$) and the explicit preference ($M = 0.49$, $SD = 0.73$, $d = 0.67$) item indicated both an explicit and implicit preference for Whites over Blacks. Participants again displayed higher levels of internal ($M = 7.44$, $SD = 1.19$) than external ($M = 5.72$, $SD = 1.48$) motivation to respond without prejudice. These

variables did not reliably differ between conditions (all F's < 0.82, all p's > 0.443).

4.2.1. IGT performance

In models analyzing each condition separately, all three conditions had a positive and reliable intercept (Black = Bad: $B = 0.29$, $t = 3.86$, $p < 0.001$, OR = 1.34, 95% CI [1.15, 1.56]; Black = Bad + Context: $B = 0.91$, $t = 6.46$, $p < 0.001$, OR = 2.48, 95% CI [1.88, 3.28]; All Black + Context = $B = 0.27$, $t = 2.97$, $p = 0.004$, OR = 1.31, 95% CI [1.09, 1.58]). However, relative to the Black = Bad + Context condition, participants had lower intercepts in the Black = Bad condition ($B = -0.54$, $t = 3.68$, $p < 0.001$, OR = 0.58, 95% CI [0.44, 0.78]) and the All Black + Context condition, ($B = -0.57$, $t = 3.93$, $p < 0.001$, OR = 0.57, 95% CI [0.43, 0.75])

For slope, in models analyzing each condition separately, all three conditions had no reliable slopes (Black = Bad: $B = 0.26$, $t = 1.88$, $p = 0.063$, OR = 1.29, 95% CI [0.99, 1.70]; Black = Bad + Context: $B = -0.01$, $t = 0.09$, $p = 0.927$, OR = 1.01, 95% CI [0.76, 1.36]; All Black + Context = $B = 0.03$, $t = 0.29$, $p = 0.771$, OR = 1.03, 95% CI [0.83, 1.29]). In addition, relative to the Black = Bad + Context condition, participants in the Black = Bad condition did not differ reliably in slope ($B = 0.21$, $t = 1.12$, $p = 0.264$, OR = 1.23, 95% CI [0.85, 1.78]), nor did participants in the All Black + Context condition, ($B = -0.003$, $t = 0.02$, $p = 0.984$, OR = 0.996, 95% CI [0.69, 1.44]).

Since no condition had a positive slope, we do not report moderator analyses here, but they are available in the online supplement.

4.3. Discussion

Moral information altered participants' ability to learn initial associations in the task. In the single-race All Black + Context condition, adding moral information created a positive intercept, as participants relied on the moral information to acquire an initial association between bad students and negative outcomes. Similarly, adding moral information to the two-race Black = Bad + Context condition created a more positive intercept, as participants used both moral and racial information to acquire the association between the Black, bad students and negative outcomes.

All conditions learned quickly to associate Black faces with negative outcomes, and adding a non-racial reason to associate Black faces with negative outcomes heightened participants' ability to learn that association initially, but participants were still unable or unwilling to strengthen that association. After being given an unambiguous, unsubtle justification to associate Black faces with negative outcomes, participants still did not strengthen anti-Black associations during the task. Though the seminary student manipulation was far from subtle, the heavy-handedness of the manipulation only reinforces the extent to which participants were unwilling or unable to strengthen an association between Black faces and negative outcomes. That is, when given a very clear manipulation that assigned Black students to a negative group of people and White students to a positive group of people, participants still did not strengthen an anti-Black association as the task progressed.

Moreover, this lack of association strengthening occurred even when other Black faces were paired with positive outcomes (the All Black + Context condition). The absence of a positive slope in a context where Black faces were associated with *both* positive and negative outcomes may suggest a more general inability or unwillingness of participants to reinforce associations that involve pairing any Black people with negative information.

5. Meta-analysis of moderators in Black = Good condition

To obtain more accurate estimates of moderators for both the intercept and slope in the Black = Good condition, we conducted an internal meta-analysis (Cumming, 2008; Lipsey & Wilson, 2001), using the MEANES SPSS macro (Wilson, 2005). We tested whether any of the following variables reliably moderated either the intercept or slope in the Black = Good condition across Studies 1–3b: IMS, EMS, explicit preferences and IAT D score, as well as task difficulty and task effort for Studies 2–3b. Analyses were run on the log-transformed odds ratio and then converted back into odds ratios for reporting. We focus only on those analyses showing a reliable moderator, but meta-analytic results for all moderators in all conditions are available in the online supplement.

Results revealed that explicit racial preferences moderated both the intercept and slope, and that IMS moderated slope. See Fig. 2 for forest plots. Higher explicit preferences for Whites relative to Blacks were associated with a lower intercept, OR = 0.83, z = −4.09, $p < 0.001$, 95% CI [0.76, 0.91], but a more positive slope, OR = 1.36, z = 3.80, $p < 0.001$, 95% CI [1.16, 1.60], meaning that participants higher in explicit preferences for Whites relative to Blacks were more likely to choose White (incorrect) faces initially, but then showed higher rates of acquiring the association between Black faces and positive outcomes, perhaps due to their lower starting point on the task.

Finally, greater IMS was associated with a more positive slope, OR = 1.23, z = 2.66, $p = 0.008$, 95% CI [1.06, 1.43], meaning that participants higher in internal motivation to control prejudice were better able to acquire the association between Black faces and positive outcomes. While we did not find reliable evidence of this IMS moderation in any individual study, running the same *Black = Good* condition across four studies allowed us to detect the effect in the internal meta-analysis.

6. General Discussion

White participants did not initially acquire an association between Black faces and positive outcomes but eventually learned this association. Conversely, White participants quickly acquired an association between Black faces and negative outcomes but did not reinforce this association. In fact, despite starting from a lower likelihood of initially selecting from a good set (a lower intercept), participants in the Black = Good condition actually ended the task with a *greater* chance of selecting a correct face from a winning set compared to participants in the Black = Bad condition.[3] In other words, by the end of the task, participants asked to learn pro-Black associations were outperforming those asked to learn anti-Black associations.

In a final study, participants still did not reinforce an anti-Black association even when given an unambiguous, non-racial reason to associate Black faces with negative outcomes. Across studies, these results align with previous work illustrating that White participants had an easier time initially learning—and a harder time unlearning—to link racial outgroups with negative outcomes (Olsson et al., 2005). Here, a more active paradigm allowed participants to control the information they sought out, revealing that Whites may learn anti-Black associations initially but resist strengthening them when they oppose prejudice motivations.

These findings depart from previous research that imposed stereotype-consistent or stereotype-inconsistent training in efforts to reduce (or increase) automatic racial bias (Johnson et al., 2016; Kawakami et al., 2000; Lai et al., 2016). Participants in our Black = Good conditions responded in ways comparable to participants in prior work asked to negate stereotype-consistent information; when given a learning context that rewards pro-Black associations, participants were willing and able to acquire and strengthen associations that countered their pre-existing anti-Black attitudes. However, participants in our Black = Bad conditions showed an interesting departure from previous studies that asked participants to negate stereotype-

[3] To complete this analysis, we recoded trial number such that the final trial had a value of 0, and then meta-analyzed differences in intercept between the two race conditions across the four studies. Participants in the Black = Good conditions had a higher likelihood of selecting from a good set on the final trial than participants in the Black = Bad conditions, OR = 1.33, $z = 3.68$, $p < 0.001$, 95% CI [1.14, 1.55]. See the online supplement for results from each study and figure.

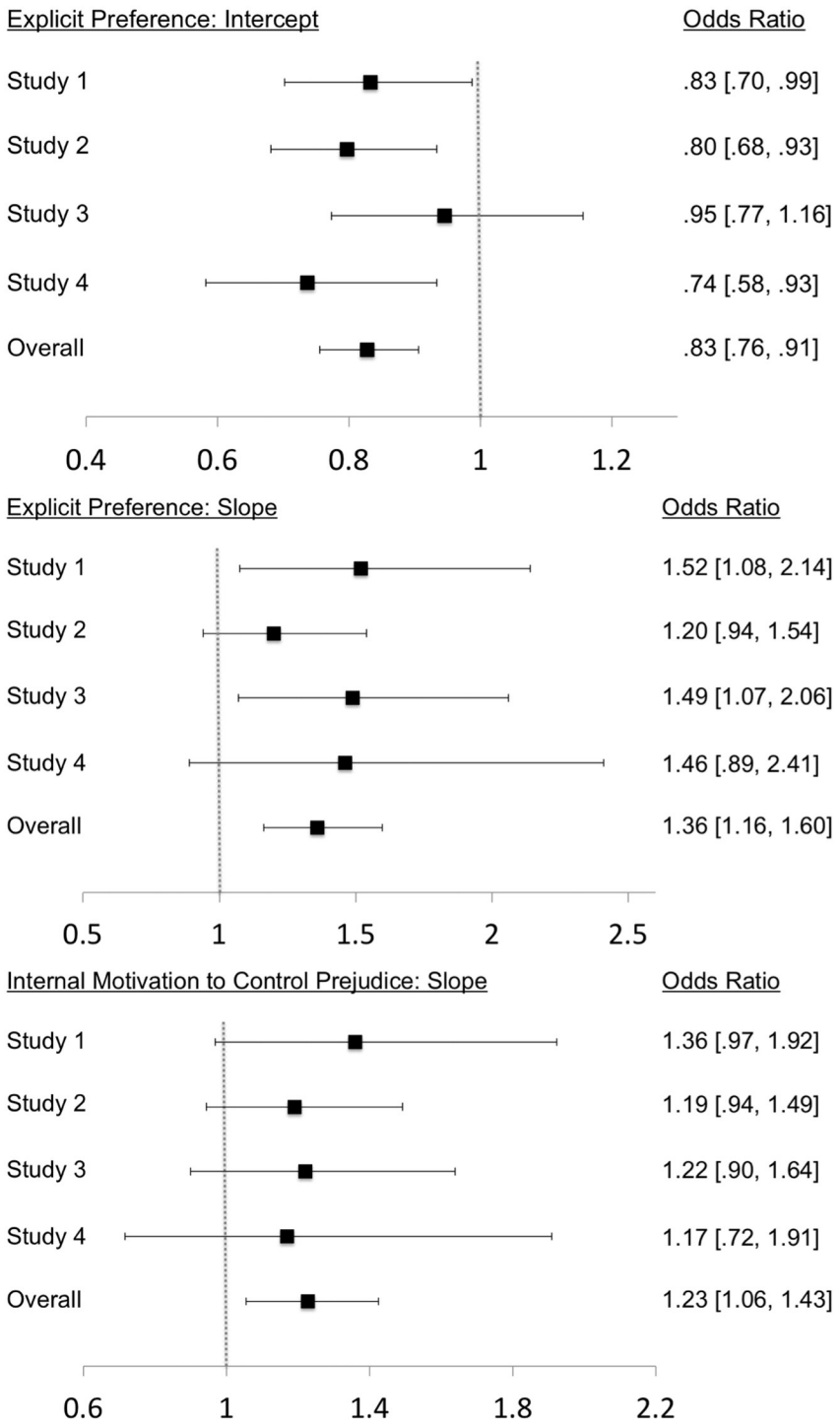

Fig. 2. Meta-analysis of moderators for slope and intercept in the *Black = Good* condition. Error bars represent 95% confidence intervals on the estimate.

inconsistent information (e.g., Johnson et al., 2016); here, when given a learning context that rewards anti-Black associations, participants were unwilling or unable to strengthen associations that align with pre-existing anti-Black attitudes. This asymmetry suggests a certain amount of control by participants over the racial associations they seek to form or strengthen that may have been overlooked by previous research.

One possibility is that these laboratory participants were altering their IGT performance to fit with perceived norms that anti-Black racial bias in behavior is unacceptable, thereby downplaying their ability to strengthen anti-Black associations simply to satisfy the experimenter's expectations. By incentivizing performance and rewarding the highest 10% of scores on the IGT, we hoped to lessen the influence of such demand characteristics, though it is unclear whether this incentive could fully remove participant concerns about expected behavior. At the very least, these studies then suggest that participants are willing to sacrifice possible monetary gain in order to not strengthen anti-Black associations that they themselves and others likely perceive as socially unacceptable.

6.1. Racial attitudes and prejudice concerns

Our results highlight the influence of both attitudes and values on behavior. On average, participants held explicit ($d = 0.75$) and implicit

($d = 1.03$) attitudes preferring Whites to Blacks, and these racial attitudes appeared to facilitate acquiring initial associations in the Black = Bad conditions, but obstructed acquiring initial associations in the Black = Good conditions. However, despite holding attitudes that favored White over Black people, most participants also valued not being or appearing racially prejudiced (e.g., an average IMS score of 7.38 out of 9 and 95% of scores above the scale midpoint), and these motivations appeared to impede strengthening associations in the Black = Bad condition but helped in acquiring associations in the Black = Good condition.

The dual role of attitudes and prejudice concerns on task performance is evident in the moderator analyses of the Black = Good conditions. Explicit racial attitudes that signaled more ingroup preference were associated with a lower intercept. Participants that more strongly preferred Whites to Blacks were more likely to begin the task by selecting White (incorrect) faces, meaning that stronger pro-White attitudes were associated with a greater difficulty in pairing Black faces with positive outcomes early in the task. However, aside from racial attitudes, race-related values and motivations also informed performance, as greater IMS was associated with a steeper slope. That is, participants more concerned about appearing unprejudiced were best at learning the association between Black faces and positive outcomes. Both racial attitudes and prejudice concerns predicted how easily participants could acquire and strengthen an association between Black faces and positive outcomes.

6.2. Unwilling or unable?

One remaining question is whether the lack of association strengthening in the Black = Bad conditions was due to conscious effort or an inability to reinforce associations between Blacks and negative outcomes. It is difficult to conclusively show whether participants were unwilling or unable to reinforce anti-Black associations, so we used both terms. However, there is some evidence that participants were consciously *choosing* not to reinforce these associations.

First, intercepts in the Black = Bad conditions were consistently above chance and slopes showed above-chance performance throughout the task, meaning participants acquired anti-Black associations quickly and maintained them. Second, the Puppies vs. Spiders condition illustrates how even associations acquired quickly can be strengthened, which did not occur in the anti-Black context of the Black = Bad condition. Third, the moderation of intercept and slope in the Black = Good condition by explicit racial preferences and IMS suggests that performance on the task in general was related to more controlled processes. Finally, our internal meta-analysis revealed that the IAT D did not reliably moderate either the intercept (Odds Ratio = 0.92, 95% CI [0.85, 1.01]) or the slope (Odds Ratio = 0.99, 95% CI [0.85, 1.14]) in the Black = Good condition, while more explicit measures (self-reported racial preference and IMS) did, indicating that IGT task performance may be more related to controlled than automatic processes. While such results are suggestive, the effort required in participants' lack of reinforcing anti-Black associations warrants further study.

7. Conclusion

On average, the participants in our studies held explicit and implicit attitudes that favored Whites over Blacks, yet simultaneously valued not being or appearing prejudiced. Here, this dissociation between race-related attitudes and prejudice concerns was clear, as both forces shaped the ability to acquire and strengthen racial associations. While attitudes are a fundamental component of behavior (Ajzen & Fishbein, 1977), this work suggests that putting more consideration into the role of prejudice concerns may help explain when behavior does not align with attitudes (e.g., Mittal & Kamakura, 2001), or even opposes attitudes (e.g., Axt, Ebersole & Nosek, 2016). Attitudes may have the first word but not the final say in learning racial information.

Appendix A

Implicit racial attitudes in all studies were assessed using a 7-block Implicit Association Test (IAT; Greenwald et al., 1998), measuring association strengths between the categories White American and Black American and the attributes Good and Bad. Participants were randomly assigned to use left or right keys for each category or attribute as well as being randomly assigned to complete White American/Good, Black American/Bad pairings first or Black American/Good, White American/Bad associations first.

In the first block (practice, 20 trials), participants categorize only images from two categories: White Americans or Black Americans using the "e" and "i" keys. In the second block (practice, 20 trials), participants categorize only words from two attributes: Good words (marvelous, superb, pleasure, beautiful, joyful, glorious, lovely, wonderful) and Bad words (tragic, horrible, agony, painful, terrible, awful, humiliate, nasty). In the third block (test, 20 trials) and fourth block (test, 40 trials) participants must categorizes both images from one category and words from one attribute jointly using the same key (e.g., images of White Americans and Good words with one key, images of Black Americans and Bad words with the other key).

In fifth block (practice, 20 trials), participants categorize only images of White Americans or Black Americans, using the opposite keys from those assigned in the first block. Finally, in the sixth (test, 20 trials) and seventh (test, 40 trials) blocks, participants categorize both images and words from one category and attribute using the same key, now completing the opposite pairing of that in the third and fourth blocks (e.g., images of Black Americans and Good words with one key, images of White Americans and Bad words with the other key).

The IAT was scored according following the guidelines of Greenwald et al. (2003) such that more positive values indicated a stronger implicit association between White American and Good and Black American and Bad. IAT scores were retained if fewer than 10% of the response trials had a latency <300ms, as recommended in Nosek et al., 2007.

References

Ajzen, I., & Fishbein, M. (1977). Attitude–behavior relations: A theoretical analysis and review of empirical research. *Psychological Bulletin*, *84*, 888–918.

Apfelbaum, E. P., Norton, M. I., & Sommers, S. R. (2012). Racial color blindness emergence, practice, and implications. *Current Directions in Psychological Science*, *21*(3), 205–209.

Axt, J. R., Ebersole, C. R., & Nosek, B. A. (2014). The rules of implicit evaluation by race, religion and age. *Psychological Science*, *25*(9), 1804–1815.

Axt, J. R., Ebersole, C. R., & Nosek, B. A. (2016a). An unintentional, robust, and replicable pro-Black bias in social judgment. *Social Cognition*, *34*(1), 1–39.

Axt, J. R., Nguyen, H. H., & Nosek, B. A. (2016b). *The judgment bias task: A reliable, flexible method for assessing individual differences in social judgment biases.* University of Virginia (in prep).

Bechara, A., Damasio, A. R., Damasio, H., & Anderson, S. W. (1994). Insensitivity to future consequences following damage to human prefrontal cortex. *Cognition*, *50*(1), 7–15.

Cumming, G. (2008). Replication and P intervals: P values predict the future only vaguely, but confidence intervals do much better. *Perspectives on Psychological Science*, *3*, 286–300.

Dunton, B. C., & Fazio, R. H. (1997). An individual difference measure of motivation to control prejudiced reactions. *Personality and Social Psychology Bulletin*, *23*(3), 316–326.

Effron, D. A., Miller, D. T., & Monin, B. (2012). Inventing racist roads not taken: The licensing effect of immoral counterfactual behaviors. *Journal of Personality and Social Psychology*, *103*(6), 916–932.

Greenwald, A. G., Klinger, M. R., & Schuh, E. S. (1995). Activation by marginally perceptible ("subliminal") stimuli: Dissociation of unconscious from conscious cognition. *Journal of Experimental Psychology: General*, *124*(1), 22–42.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology*, *74*(6), 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*(2), 197–216.

Johnson, I. R., Kopp, B. M., & Petty, R. E. (2016). Just say no!(and mean it): Meaningful negation as a tool to modify automatic racial attitudes. *Group Processes & Intergroup Relations*. http://dx.doi.org/10.1177/1368430216647189 (Advance online publication).

Kawakami, K., Dovidio, J. F., Moll, J., Hermsen, S., & Russin, A. (2000). Just say no (to stereotyping): Effects of training in the negation of stereotypic associations on stereotype activation. *Journal of Personality and Social Psychology*, *78*(5), 871–888.

Kline, R. (2011). *Principles and practice of structural equation modeling* (3rd ed.). New York, NY: Guilford Press.

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.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage.

Merritt, A. C., Effron, D. A., Fein, S., Savitsky, K. K., Tuller, D. M., & Monin, B. (2012). The strategic pursuit of moral credentials. *Journal of Experimental Social Psychology*, *48*(3), 774–777.

Mittal, V., & Kamakura, W. A. (2001). Satisfaction, repurchase intent, and repurchase behavior: Investigating the moderating effect of customer characteristics. *Journal of Marketing Research*, *38*(1), 131–142.

Monin, B., & Miller, D. T. (2001). Moral credentials and the expression of prejudice. *Journal of Personality and Social Psychology*, *81*(1), 33–43.

Nosek, B. A., Banaji, M. R., & Greenwald, A. G. (2002). Harvesting implicit group attitudes and beliefs from a demonstration web site. *Group Dynamics: Theory, Research, and Practice*, *6*(1), 101–115.

Nosek, B. A., Greenwald, A. G., & Banaji, M. R. (2007). The implicit association test at age 7: A methodological and conceptual review. *Automatic processes in social thinking and behavior* (pp. 265–292).

Olsson, A., Ebert, J. P., Banaji, M. R., & Phelps, E. A. (2005). The role of social groups in the persistence of learned fear. *Science*, *309*(5735), 785–787.

Plant, E. A., & Devine, P. G. (1998). Internal and external motivation to respond without prejudice. *Journal of Personality and Social Psychology*, *75*(3), 811–832.

Trawalter, S., & Richeson, J. A. (2008). Let's talk about race, baby! When Whites' and Blacks' interracial contact experiences diverge. *Journal of Experimental Social Psychology*, *44*(4), 1214–1217.

Wilson, D. B. (2005). Meta-analysis macros for SAS, SPSS, and Stata. (Retrieved March, 14, 2016 from) http://mason.gmu.edu/~dwilsonb/ma.html.