DRAFT: Published as Fritzlen, K. A., Phillips, J. E., March, D. S., Grzanka, P. R., & Olson, M. A., (2019). I know (what) you are, but what am I? The effect of recategorization threat and perceived immutability on prejudice. ***Personality and Social Psychology Bulletin***

I know (what) you are, but what am I?: The effect of recategorization threat and perceived immutability on prejudice

Katherine A. Fritzlen, Joy E. Phillips, David S. March, Patrick R. Grzanka, Michael A. Olson

University of Tennessee

Word Count (with Abstract): 10,365

Abstract

Learning one is similar to a stigmatized group can threaten one's identity and prompt disassociation from the group. What are the consequences of learning of a similarity to a stigmatized group when that similarity implies possible recategorization into the group? We investigated how learning of an immutable, recategorization-implying similarity with an outgroup affects implicitly- and explicitly-assessed prejudice. In Study 1, White participants who believed they had above average genetic overlap with African Americans showed decreased prejudice on implicit but not explicit measures. In Study 2, straight/heterosexual participants who were led to believe they exhibited some same-sex attraction showed reduced implicitly-assessed prejudice, but only if they believed sexual orientation was biologically-determined. Thus, learning of an identity-implying similarity with an outgroup can reduce implicit prejudice if that group membership is believed to be immutable. Theoretical and practical implications are discussed.

KEYWORDS: prejudice, social identity, intergroup processes, psychological essentialism

I know (what) you are, but what am I?: The effect of recategorization threat and perceived immutability on prejudice

Contemporary biomedical technology makes it is easy for people to uncover their genetic heritage. But not everyone receives the results they expect. We wondered whether different processes might manifest implicitly and explicitly when an individual is confronted with evidence suggesting they share essential, group-defining similarities with a stigmatized outgroup. Specifically, does confrontation with the possibility of membership in a marginalized outgroup have the potential to *reduce* prejudice at the implicit level?

Little research has addressed this question, and what has lacks experimental controls. For example, many White supremacists use genetic tests to prove their "pure" Aryan ancestry. Panofsky and Donovan (2017) analyzed thousands of online posts of White supremacists' explicit reactions to genetic tests indicating some non-European ancestry. These individuals engaged in a variety of strategies, including rejecting the results, offering justifications, reinterpreting history, and questioning the standards by which ancestry was determined. A few even appeared to accept the results and rethink their white nationalist beliefs. Thus, at least explicitly, some individuals, when confronted with evidence that challenged their preferred identity, distance themselves from an undesirable social identity. Yet, it remains to be seen whether the same process would play out implicitly. This curiosity sparked the current work.

We induced White participants to believe they showed above average genetic similarities to Black Americans (Experiment 1), and straight participants to believe they showed sexual arousal patterns similar to gay or lesbian individuals (Experiment 2). In both experiments, we predicted individuals who received feedback implying membership in a stigmatized outgroup

would exhibit increased explicitly-assessed prejudice. However, we also expected a concomitant decrease in implicitly-assessed prejudice towards that group.

Social Identity Threat

Social Identity Theory (Taijfel & Turner, 1979) asserts that people define themselves and derive a sense of worth from the characteristics and achievements of the groups with which they affiliate (Turner, 1999). Belonging to a group fosters camaraderie and closeness and can enhance self-esteem when the group possesses positive characteristics. Indeed people's evaluations of groups they belong to are associated with their self-evaluations (Greenwald, Banaji, Rudman, Farnham, Nosek, & Mellott, 2002). Since one's group-evaluations are closely tied with one's self-evaluations, people are motivated to associate with groups that possess positive characteristics.

Conversely, possible membership in a disliked group is threatening and can cause disassociation from (Cialdini, Borden, Thorne, Walker, Freeman, & Sloan, 1976; Novak & Lerner, 1968; Schimel, Pyszczynski, Greenberg, O'Mahen, & Arndt, 2000), and increased prejudice and discrimination towards that group (Adams, Wright, & Lohr, 1996; Gibbons, 1985; Zarate, Garcia, Garza, & Hitlan, 2004). For example, when induced to think of how they were similar to immigrants, Zarate et al.'s participants showed greater explicit prejudice toward immigrants than when they thought about how they were different from them. In short, perceiving oneself as similar to a negatively-stereotyped group is a threat to one's social identity.

Previous research investigating similarities to outgroups has investigated group-*relevant* but not group-*defining* traits. That is, none of the previous research has investigated the impact of similarities that have the power to determine *identity* with an outgroup. In contrast to previous

research, the current investigation focuses on perceptions of a similarity on a *group-defining* trait, one that implies a new group identity.

Categorization Threat

Branscombe, Ellemers, Spears, and Doosje (1999) outlined a taxonomy of four types of social identity threats (i.e., threats of: categorization, distinctiveness, value, and acceptance). Categorization threat, the threat of being categorized against one's will, is most closely related to the identity threat of group-defining similarities manipulated in the current investigations. When a group's attributes are positive, an individual typically welcomes acceptance into a group. However, if the group is perceived negatively, placement into that group can result in a categorization threat. In other words, this threat arises when people's preferred self-categorizations do not correspond to the way they are perceived by others. At least at the explicit level, this can lead to defensive reactions (Long & Spears, 1997), including dissociation from the new group (Ellemers, Wilke, & Van Knippenberg, 1993) and derogation of group members (Meindl & Lerner, 1984), particularly when the group is stigmatized.

Likewise, in the current research, the threat we manipulate does not arise from making a negatively characterized social identity salient; rather, we explore the threat of being potentially *recategorized* into a devalued social identity group. Like the threat experienced by White supremacists whose genetic ancestries were revealed to be less than 100% European, participants in the current investigation experienced a similar recategorization threat when they received feedback that called their racial (Study 1) or sexual identities (Study 2) into question.

Implicit vs. Explicit Effects of Categorization Threat

We argue the categorization threat to our participants will decrease implicitly-assessed but not explicitly-assessed prejudice. Implicit measures can reveal attitudes that explicit

measures cannot because the latter are subject to normative pressure (Fazio, Jackson, Dunton, & Williams, 1995) and more careful, reasoned responding (Hofmann, Gawronkski, Gschwendner, Le, & Schmitt, 2005). When the domain in question is not socially sensitive (e.g., Nosek, Banaji, & Greenwald, 2002), implicit and explicit measures generally align.

Theoretical models of evaluation like the APE model (Gawronski & Bodenhausen, 2006) and MODE model (Fazio & Olson, 2014) argue that spontaneous, affective responses captured by implicit measures are largely associative in nature. Such associations can form when a neutral object is paired with a positive or negative object, creating an evaluative association (Olson & Fazio, 2001). These associations can be automatically activated upon perception of the object, which implicit measures are well-suited to capture, but may or may not influence downstream, deliberative judgments on explicit measures, depending on other activated associations (Gawronski & Bodenhausen, 2006) or relevant motives (Fazio & Olson, 2014). For example, Olson and Fazio (2006) paired Black faces with positive objects and White faces with negative objects, and this associative learning resulted in reduced prejudice on implicit but not explicit measures. These authors argued that normative beliefs and motivated processes (such as the desire to avoid the appearance of prejudice) influenced responses on explicit measures.

We argue that categorization threat fosters an association between the self and the newly joined ingroup (i.e., the previous outgroup), resulting in increased positive associations regarding the group. Previous research has found that people tend to evaluate things associated with the self positively (Gawronski, Bodenhausen, & Becker, 2007; Greenwald et al., 2002). For example, people prefer letters of the alphabet in their own name (Nuttin, 1985), evaluate objects more favorable merely because they own them (Beggan, 1992), and express more liking for just-chosen items than items not chosen (Brehm, 1956).

We know that people generally evaluate the groups they are a part of more positively than groups they are not both explicitly (Tajfel, Billig, Bundy, & Flament, 1971) and implicitly (Perdue, Dovidio, Gurtman, & Tyler, 1990). Like other objects, merely being categorized into a new group may be sufficient for an associative transfer to occur. Research on the minimal groups paradigm, for example, has found that when individuals are randomly sorted into a laboratory-created groups, they exhibit automatic preference for the artificial ingroup (Ashburn-Nardo, Voils, & Monteith, 2001; Otten & Wentura, 1999). Thus, the process through which people develop more positive evaluative associations toward groups associated with the self can occur relatively quickly.

We suspect that perceiving oneself as recategorized into a group (even a negatively-characterized one) should facilitate an association between the group and the self, leading to an associative transfer of the one's implicit self-evaluations to the group (Walther & Trasselli, 2003). Given that most people have positive implicit self-evaluations (Koole, Dijksterhuis, & van Knippenberg, 2001), associative self-anchoring may lead to enhancement of implicit evaluations of the new ingroup. Yet, as the new ingroup is negative, and membership is presumably not desired, explicit measures may reflect deliberate social distancing and other downstream motivations (according to dual-process models), obscuring any newly-formed associations.

It is important to note that some dual-systems theories (e.g., Rydell, McConnell, Strain, Claypool, & Hugenberg, 2007) argue that automatic evaluations are output of a slow-learning associative system that requires many repetitions of new information to produce change in evaluations, whereas deliberate evaluations reflect the output of a relatively fast-learning system that can change based on a single input. Our argument—that a single piece of identity-

threatening information changed implicitly-assessed but not explicitly-assessed attitudes—would appear to be at odds with such models. However, recent theory and research challenge the assumption that automatic evaluative associations are slow to change (e.g., Brannon & Gawronski, 2017). In fact, the APE model (Gawronski & Bodenhausen, 2006) and single-process propositional models of evaluative learning (De Houwer, 2014) identify pathways to automatic evaluative association change on the basis of a single piece of information (i.e., in the absence of repeated exposure or rehearsal of that information). Mann and Ferguson (2015) provide insight into the conditions under which such change occurs, namely, when the new information prompts reinterpretation of previously learned information, and when time and cognitive resources are available. Novel information must also be perceived as highly diagnostic to reverse automatic evaluations (Cone & Ferguson, 2015; see also Brannon & Gawronski, 2017).

Thus, we suggest that the self-association to the outgroup formed on the basis of an identity threat likely emerges through deliberation on the newly learned information about the self: one must reconcile one's (presumably) positive self-views with inclusion in an outgroup, resulting in an automatic evaluation of the outgroup that is more positive. However, as discussed above, because explicit measures allow for input of more motivated processes and self-presentational concerns, we considered it unlikely that participants would be comfortable expressing explicitly any reductions in prejudice on the basis of the shift in self-identity.

Current Work

The goal of the present research is to explore how learning of an immutable similarity, i.e., possible membership with a stigmatized outgroup, affects attitudes towards that group. Study 1 investigated the effect of learning about genetic similarity to Black individuals on White

individuals' implicit and explicit prejudice. We predicted individuals given feedback implying a biological similarity with an outgroup would exhibit reduced implicit prejudice, but not explicit prejudice. Study 2 explored the same functional effect and examined essentialist beliefs regarding the biological basis of the category as a potential moderator.

Study 1

Methods

Participants and Design. Eighty-eight White undergraduate psychology students at a large southeastern U.S. university were recruited via a web-based system and participated for partial fulfillment of course requirements. Participants completed one 15-minute group lab session (Session 1), after which they signed up for a second 30-minute individual follow-up session (Session 2). Six participants were removed from analyses for high error rates (> 25%) on the Implicit Association Test (IAT), resulting in a final sample of 82 participants who were randomly assigned to one of two feedback conditions: White feedback condition (n = 40) or more-Black feedback condition (n = 42). We made no theoretically-informed estimates of effect size, so we collected as many participants as we could given the resources available.

Pretest. At the beginning of the semester, participants completed an online prescreening survey in which they provided demographic information. Only participants who self-identified as White were able to view and then register for the experiment.

Session 1. Upon arriving to the lab in groups of 1-5, participants were greeted by a White experimenter who told them the experiment was about group membership and that they would be submitting cheek swab samples of their DNA to accurately determine their group membership. Participants were told that, in order to classify them, genetic markers from their DNA sample

would be analyzed and compared to samples from other groups. In reality, this "sample" was discarded. The experimenter then assisted each participant in obtaining their sample. To bolster the cover story, the experimenter wore white lab coats and exam gloves, used medical swabs to take the samples, and stored the samples in individual bags emblazoned with medical biohazard text and symbols. Following the "DNA collection", each participant registered for a follow-up session to receive their results and complete the dependent measures.

Session 2. Participants completed Session 2 individually two days after Session 1. When they arrived, they were escorted to a private room by a White experimenter where they received their DNA results. We expected that White participants would not believe being told that they were "Black"; instead, we manipulated the extent to which participants' feedback implied that they might have a recent Black ancestor. In other words, we induced them to believe they were "more Black" (and consequently less White) than they may have previously believed.

Regardless of condition, each participant received a handout indicating, "This sample shares 53% genetic overlap with African Americans." While all participants received this same feedback about the overall genetic overlap, the "average" genetic overlap between White and African Americans was manipulated between conditions. Specifically, the handout in the White feedback condition continued: "On average, most Caucasian individuals share 75% [overlap with African-Americans]. This indicates that the individual from whom this sample was taken has less than the expected genetic overlap with African-Americans. What these results mean is that you have a lower than average amount of genetic overlap with African-Americans. Most individuals who aren't African-American share about 75% genetic overlap with them, but it looks as though you only share 53%. This indicates that the number of gene sequences you share with African-Americans is a little less than average."

Those in the more-Black feedback condition were provided a different frame of reference in their feedback. Specifically, those in the more-Black feedback were told: “On average, most Caucasian individuals share 10% [overlap with African-Americans]. This indicates that the individual from whom this sample was taken has more than the expected genetic overlap with African-Americans. What these results mean is that you have a higher than average amount of genetic overlap with African-Americans. Most individuals who aren’t African-American only share about 10% genetic overlap with them, but it looks as though you share 53%. This doesn’t necessarily mean that you have African-American ancestors in your *immediate* family, but it does indicate that a number of your gene sequences are highly similar to African-Americans, and it’s likely that you have at least one ancestor who is African-American.” Following this report, the experimenter answered any questions participants had about their results. No participant expressed doubt about the feedback.

Dependent measures. Next, all participants completed a race IAT, feeling thermometer, and trait ratings of themselves, individually, and Blacks, as a group, in that order. We used a personalized five-block IAT with 30 trials per practice block and 60 trials per critical block (Olson & Fazio, 2004). The third and fifth blocks contained critical trials in which participants used the same response key to categorize both a particular race and the valence of an adjective. The order of exposure to critical blocks was counterbalanced across subjects and reaction times for responses on the critical blocks were recorded using DirectRT (Jarvis, 2014). IAT block order was a between-participants counterbalancing manipulation that yielded no effects. We computed IAT d-scores according to recommendations by Greenwald, Nosek, and Banaji (2003), such that higher numbers indicated a pro-White attitude. However, error trials were removed

from analyses in lieu of an error penalty, along with any trial with raw latencies shorter than 150 milliseconds (ms) or longer than 2500 ms, as in Olson and Fazio (2004).

The feeling thermometer prompted them to rate how positively or negatively they felt towards a variety of social groups from 0 (very cold) to 100 (very warm). Social groups were presented in random order and included Black people, White people and filler social groups (e.g., Republicans, Hispanics). We computed a standardized feeling thermometer index of explicit prejudice toward Blacks for each participant by subtracting the mean rating of all the other social groups from the rating of Blacks, then dividing the difference by the standard deviation of the ratings of all social groups. More positive numbers indicated more positive attitudes of Blacks relative to other social groups.

Finally, participants made trait ratings of themselves and Black people, respectively (traits were taken from Wolsko, Park, Judd, & Wittenbrink, 2000). Half the items were stereotypic of Whites and counterstereotypic of Blacks, and half were stereotypic of Blacks and counterstereotypic of Whites; additionally, half had a positive valence, and half had a negative valence, for a total of four trait types. Participants rated 56 items from 0 (does not describe at all) to 100 (describes very well) for both themselves and Blacks. To assess the possibility that participants might dissociate themselves from Blacks regardless of the direction (positive or negative) of distancing, mean overall distancing scores were calculated for each participant using the absolute value of the difference between participants' ratings of themselves and Black Americans on all 56 traits (based on the procedures used by Schimel et al., 2000) with higher scores indicating greater distancing. After the trait ratings, participants were asked some final

open-ended questions about the experiment (where none expressed suspicion), fully debriefed, thanked, and dismissed.[1]

Results

Preliminary Analyses. Results from a one-sample t-rest showed an overall anti-Black prejudice effect on IAT d-scores, $M = .40$, $SD = .36$, $t(81) = 10.16$, $p < .001$, $d = 2.26$. Results from a one-sample t-test for feeling thermometer scores showed no overall anti-Black prejudice effect, $M = -.12$, $SD = .30$, $t(81) < 1$. Other ways of computing the thermometer (e.g., by creating a difference score) did not produce any unique effects. For trait ratings, results from one-sample t-tests showed an overall distancing effect on the trait ratings, $M = 1.35$, $SD = .25$, $t(81) = 48.52$, $p < .001$, $d = 10.78$. Other ways of creating distancing scores (e.g., by accounting for trait valence) yielded no unique effects. Table 1 reports correlations between these measures.

Effect of Feedback Condition on Prejudice. To assess the effects of feedback condition on implicit and explicit prejudice, we conducted independent samples t-tests on IAT d-scores, feeling thermometer *z*-scores, and the trait rating absolute difference scores. For the IAT, we observed the expected main effect of feedback condition, $t(80) = 2.71$, $p = .008$, $d = 0.62$, such that participants in the more-Black feedback condition showed lower prejudice ($M = .30$, $SD = .33$, 95% CI [.20, .40]) than those in the White feedback condition, $M = .51$, $SD = .35$, 95% CI [.39, .62]; see Figure 1. That is, being told that they were more genetically similar to Blacks than the typical White person reduced participants' implicit prejudice. There was no difference between the more-Black and White feedback conditions for the feeling thermometer or trait ratings, all $|t|$'s < .92, all p's > .3. In other words, receiving White compared to more-Black feedback did not differentially affect participants' explicit prejudice.

[1] The local institutional review board reviewed and approved both studies and no participants exhibited or reported distress due to their participation in either study.

Discussion

When White participants were told they showed above-average genetic similarity to African-Americans, they showed a reduction in implicit prejudice, with no corresponding change in their explicit prejudice. However, it is unclear whether this change in attitudes is unique to race, which people tend to believe is immutable (Haslam, Rothschild, & Ernst, 2000), or whether a similar change in attitudes would be seen for other social categories.

As a social-cognitive process, categorization threat is arguably most impactful when it invokes individuals' perceptions of social groupings as reflecting a fundamental essence, particularly one that is immutable and innate. Indeed, if group membership was mutable, one could simply renounce one's membership when threatened with categorization in a devalued outgroup. On the other hand, if membership in a given group is not seen as mutable, but rather fixed and stable, individuals may not have the option to engage in such exit strategies.

Social categories vary in the extent to which people believe membership is mutable. While most White Americans believe that race is biologically-based and immutable, people do not hold such uniform beliefs about the mutability of sexual orientation (Haslam et al., 2000). For example, while many people believe that sexual orientation is biologically determined (Grzanka, Zeiders, & Miles, 2016), there are programs (e.g., so-called "conversion" therapies) based on the idea that it is possible to change one's sexual orientation if one is motivated and willing (Waidzunas, 2015). It may be that the attitudinal change found in Study 1 would only occur for groups where membership is seen as immutable. Study 2 investigated this possibility.

Study 2

The immutability of social categories is part of a larger set of beliefs known as psychological essentialism. Psychological essentialism is the process by which individuals come

to believe that social groupings are natural, basic, non-arbitrary, and therefore, indicative of some kind of essence, as opposed to socially constructed (Medin, 1989). Yzerbyt, Rocher, and Schadron (1997) developed a taxonomy of essentialist beliefs, including beliefs that social categories are discrete, uniform, informative, natural, immutable, stabile, inheritable, necessary, and exclusive. They argue that essentialist beliefs can serve to rationalize and legitimize stereotypes and treatment of stigmatized groups. Haslam and colleagues (2000) assert that psychological essentialism facilitates the process by which members of social groups come to be viewed as linked through a fundamental property, as well as assigned a kind of fixity and uniformity; these labels then facilitate inferences about virtually all members of a social group.

A robust body of empirical research has demonstrated a fairly consistent relationship between essentialist beliefs about race and gender and prejudicial attitudes toward women and racial minorities (e.g., Haslam, Rothschild, & Ernst, 2002). For example, the belief that African Americans are more similar to one another than they are to White people buttresses racism; similarly, endorsement of the idea that women are fundamentally less capable of certain intellectual tasks than men reinforces sexist ideology.

When it comes to sexual orientation categories, however, the intersection of attitudes and lay beliefs exposes a complex relationship between essentialism and prejudice. Haslam and Levy (2006) and Hegarty and Pratto (2001) showed how some essentialist beliefs, such as belief in the naturalness and immutability of sexual orientation, corresponded with positive attitudes toward sexual minorities, whereas belief in the discreteness of sexual orientation groups did not. In fact, Hegarty and Pratto (2001) found that greater endorsement of naturalness beliefs was associated with more pro-gay attitudes. Meanwhile, among college students, Grzanka and colleagues (2016) found that naturalness beliefs—or the idea that sexual minorities are “born this way”—did not

distinguish those who held positive or negative attitudes toward gay men as much as other belief domains did (e.g., discreteness of sexual orientation categories and the homogeneity of sexual orientation group members).

We were interested in whether the effect of our categorization threat would differentially affect prejudice as a function of individual' beliefs in the immutability of sexual orientation. Seeing as naturalness taps into beliefs about the fixedness and immutability of social categories, the concept of naturalness was of particular interest in Study 2.

Overview

The first goal of Study 2 was to conceptually replicate the findings of Study 1 with a different social identity. Specifically, we investigated how straight individuals' attitudes towards gay individuals would be affected by being led to believe that their sexual preferences were more similar to gay men and lesbian women than the average straight person. Secondly, we sought to examine the potential moderating effects of beliefs regarding the naturalness (i.e., innateness and immutability) of the identity.

Consistent with our findings in Study 1, we predicted that categorization threat would result in reduced implicit prejudice, but beliefs in the biological immutability of the social category would moderate this effect. Specifically, we expected straight participants told they exhibited relatively stronger patterns of same-sex attraction than the average straight individual would exhibit reduced implicit prejudice toward gay men and lesbian women, but only to the extent that they endorsed naturalness beliefs about sexual orientation. While we initially expected to see explicit distancing as a function of categorization threat, we were hesitant to make this prediction in Study 2 given that we saw no such pattern in Study 1.

Methods

Participants and Design. Undergraduate students (*n* =191) were recruited via a web-based sign-up system through a large Southeastern university and participated in exchange for course credit. Participants who identified as gay or lesbian were excluded from the analyses (*n* = 2). Thirty-three participants failed to complete some or all of the moderating (*n* = 18) or explicit measures (*n* = 15). Analyses are reported only for participants with full data sets, but the pattern of effects were unaltered by including these participants in analyses. Our final dataset included one hundred and fifty-six participants (72 Women, 82 Men, and 2 who identified as neither, whose inclusion also did not alter the pattern of results) who were randomly assigned to one of two conditions: relatively more gay feedback (*n* = 82) or straight feedback (*n* = 74). Based on the results of Study 1, in order to detect an effect size of .62 with 80% power, we needed 84 participants, which our sample size exceeded.

Pretest. At the beginning of the semester, participants completed an online prescreening survey in which they provided demographic information and essentialist beliefs about sexual orientation using the Sexual Orientation Beliefs Scale (SOBS). Developed by Arseneau, Grzanka, Miles, and Fassinger (2013), the SOBS measures a range of essentialist-related beliefs about sexual orientation. This scale consists of 35 items along 4 distinct factors: naturalness, discreetness, entitativity, and importance. Participants rated each of the statements on a 7-point Likert scale ranging from -3 (strongly disagree) to +3 (strongly agree; note: the original published scale employed 5-point response options). The naturalness subscale, the focus for the present investigation, consisted of 11 items that assessed one's belief that sexual orientation (SO) is innate, biologically-based, immutable, has early fixity, and is stable across cultures (e.g., "Biology is the main basis of an individual's sexual orientation." and "If someone comes out as gay or lesbian they were probably attracted to the same sex all along."). The mean score for the

naturalness subscale of the SOBS was computed such that higher scores indicated higher belief support (α = .72). This score was treated as a continuous variable in all analyses.

Laboratory Procedure. Participants were tested in groups of 1-5. Upon arriving at the lab, they were greeted by an experimenter and seated at individual computers with privatizing dividers. Participants were led to believe that they would be completing a study on impression formation and word processing. All tasks were administered using MediaLab and Direct RT software (Jarvis, 2014).

First, all participants completed the 10-item Rosenberg Self-Esteem Scale (Rosenberg, 1979) for exploratory purposes (α = .74). Next, all participants completed an "Impression Formation Task" in which they were shown images of individuals and told to form an impression of each, as they would purportedly be asked about them later. Specifically, participants saw 10 men and 10 women, some of whom wore revealing clothing, for 8 seconds each.

After participants completed this task they were informed that the task was actually examining pupil dilation in response to arousing stimuli. In actuality, the eye trackers underneath the monitor were fake, but they were not made aware of this until they completed the experiment. No participants expressed doubts about the authenticity of the ostensible eye-trackers.

Just as we did not think it would be believable to give White participants feedback that they were actually Black, we did not think it would be believable to give our largely self-reportedly straight population feedback implying they were actually gay. Instead, the more-gay feedback implied that their sexual arousal patterns (indicated by pupil dilation) were relatively more similar to gay individuals than the typical straight person or, in other words, that they were

"more gay" than they may have previously believed. The straight feedback condition implied dilation patterns very closely resembling the typical heterosexual individual.

Participants were shown ostensible images of the size of their pupil dilation in response to same-gender and opposite-gender images next to the average size of gay and straight participant pupil dilations to the same images (Figures 2.1 and 2.3). Participants in the straight feedback condition were presented with images of pupil dilation patterns that were practically identical to the average straight participant's pupil dilation patterns, implying typical heterosexual patterns of arousal (Figures 2.1 and 2.2). Participants in the more-gay feedback condition were shown that their pupil dilations were slightly smaller to opposite-sex images and slightly larger for same sex images than the average straight participants', implying that they showed more sexual arousal to same-sex images and less sexual arousal to opposite-sex images than the average straight person (Figures 2.3 and 2.4). On the next screen, participants were shown the percentage that their dilations were larger or smaller than the average straight individual's pupil dilations to same- and opposite-sex images. Those in the straight feedback condition were told their pupil dilations were 7% larger when looking at opposite sex images and 0% larger when looking at same sex images (Figure 2.2). Those in the more-gay feedback condition were told their pupil dilations were 11% smaller when looking at opposite sex images and 28% larger when looking at same sex images (Figure 2.4).

On the next screen, all participants were given feedback on the implications of their results. Those in the straight feedback condition were told: "This implies that your sexual attraction patterns are very similar to the average heterosexual participant, and very different

from the average homosexual participant."[2] Those in the more-gay feedback condition were told: "This implies that your sexual attraction patterns differ from the average heterosexual participant, and showed some similarity to the average homosexual participant." All participants then rested for 30 seconds before the next task.

Dependent Measures. Next participants completed an IAT analogous to that from Study 1. The 4 categories of objects were pleasant words, unpleasant words, and romantic pictures of straight couples and gay couples. The image stimuli were obtained through Google image search and were labeled "free to use or share." IAT block order was a between-participants counterbalancing manipulation that yielded no effects. We computed IAT d-scores according to the same procedure as Study 1.

To assess explicit prejudice, participants completed a feelings thermometer (Olson & Zabel, 2016), made trait ratings of themselves, gay men, and lesbian women. The feelings thermometer was identical to that of Study 1, including the target groups (gay men and lesbian women) as well as several filler groups. Standardized feeling thermometer scores were computed as in Study 1.

For the trait ratings, participants were asked to rate how well a series of 30 different traits described them from 0 (does not describe at all) to 100 (describes very well) for themselves, gay men, and lesbian women, separately. Many, but not all, of the traits were drawn or adapted from Bem's Sex Role Inventory (Bem, 1977) and included stereotypically feminine traits (e.g., compassionate, feminine, loyal), masculine traits (e.g., forceful, competitive, assertive), as well as androgynous traits (e.g., secretive, happy, unpredictable). To examine the extent to which

[2] We acknowledge that "homosexual" is neither person-first language nor consistent with APA Style. We used the term in the manipulation only, so that feedback could be rhetorically consistent across conditions. In this manuscript, we use person-first language whenever possible.

participants distanced themselves explicitly, absolute value difference scores of the trait ratings were created just as in Study 1, but separately for gay men and lesbian women.

After the trait ratings, participants completed the 24-item Modern Homonegativity Scale (Morrison & Morrison, 2003; $\alpha = .97$). We included this measure to address questions beyond the scope of the present investigation and thus it will not be discussed further. Participants were then asked some final open-ended questions about the experiment (where none expressed any suspicion or doubt about the feedback manipulation), were fully debriefed, offered an opportunity to revoke their consent (none refused) thanked, and dismissed.

Results and Discussion

Preliminary Analyses. Results from a one-sample t-rest showed an overall anti-gay prejudice effect on the IAT, $M = .63$, $SD = .44$, $t(155) = 17.70$, $p < .001$, $d = 2.84$. For standardized feeling thermometer scores, results from one-sample t-tests showed an overall prejudice effect on the feelings thermometer ratings for gay men, $M = -.40$, $SD = 1.163$, $t(155) = -4.33$, $p < .001$, $d = 0.70$, and lesbian women, $M = -.33$, $SD = 1.07$, $t(155) = -3.89$, $p < .001$, $d = 0.62$. Results from a one-samples t-test looking at the extent to which participants distanced themselves explicitly showed an overall distancing effect on the trait ratings regarding self and gay men, $M = .95$, $SD = .27$, $t(155) = 43.98$, $p < .001$, $d = 7.07$, as well as self and lesbian women, $M = .98$, $SD = .25$, $t(155) = 48.73$, $p < .001$, $d = 7.83$. As in Study 1, various other ways of creating distancing scores yielded no unique effects. Overall, scores on the naturalness beliefs subscale did not significantly vary by condition, $t(154) = 1.161$, $p = .25$. See Table 2 for correlations between dependent measures.

Impact of Naturalness Beliefs and Feedback Condition on Prejudice. To assess the effects of naturalness beliefs and feedback condition on implicit and explicit prejudice, in the

following analyses we separately regressed the IAT, feeling thermometer *z*-scores (separately for gay men and lesbians), and trait rating absolute difference score measures (separately for gay men and lesbians) on feedback condition and the naturalness subscale of the SOBS. On the first step of the regression, condition assignment (dummy-coded) and main effects of naturalness were entered. Their interaction term was entered on step 2.

Implicit Prejudice. For IAT scores, there was neither a main effect of naturalness, b = -.04, SE = .04, $t(153) = -1.00$, p = .32, nor a main effect of feedback condition, b = -.04, SE = .07, $t(153) = -.56$, p = .58. However, the predicted Naturalness X Feedback interaction was significant, b = .23, SE = .08, $t(152) = 2.83$, p = .005, 95% CI [.07, .40]. Further analyses by condition revealed that in the more-gay feedback condition, higher naturalness scores were associated with reduced prejudice on the IAT, as expected, b = -.16, SE = .06, $t(80) = -2.67$, p = .01, 95% CI [-.28, -.04]. There was no relationship between naturalness and IAT scores in the straight feedback condition, b = .07, SE = .06, $t(72) = 1.30$, p = .20, 95% CI [-.04, .19] (see Figure 3).[3]

Explicit Prejudice. For feelings thermometer ratings, there was a significant effect of naturalness on participants' ratings of gay men such that participants with higher naturalness beliefs expressed warmer attitudes towards gay men, b = .28, SE = .11, $t(153) = 2.62$, p = .01.

[3] Prior to analyses, the first and fifth author determined that participants who self-identified as gay or lesbian would be excluded from analyses, but did not make a priori exclusion decisions about those identifying as bisexual and/or "other" for sexual orientation or gender. Thus, participants identifying as such were included in the analyses reported. However, such a decision fails to acknowledge the unique identities and experiences of bisexual, transgender, and gender nonconforming individuals. Thus, we conducted additional analyses that only included participants who self-identified as straight and either man or woman, i.e., cisgender, heterosexual individuals. The feedback condition x naturalness interaction was still significant, $t(147) = 2.53$, p = .013, and the relationship between Naturalness and Implicit prejudice was still observed, albeit slightly weaker for the more-gay feedback condition, $t(76) = -1.95$, p = .055, B = -.22, and slightly stronger for the straight feedback condition, $t(76) = 1.61$, p = .111, B = .19.

There was neither a main effect of condition nor an interaction between naturalness and condition for thermometer ratings of gay men, all *t*'s < 1.4, all *p*'s > .1. There were no main effects of naturalness, condition, or their interaction found for thermometer ratings of lesbians, all *t*'s < 1.2, all *p*'s > .2.

For the explicit trait ratings, there was a marginally significant effect of feedback condition on participants trait ratings of gay men, *b* = -.08, *SE* = .04, *t*(153) = -1.83, *p* = .07. Those in the more-gay feedback condition showed marginally increased self-gay differences in trait ratings, *M* = .99, *SD* = .27, 95% CI [.93, 1.05], compared to those in the straight feedback condition, *M* = .91, *SD* = .26, 95% CI [.85, .97]. For trait ratings of lesbians, there was a non-significant trend such that higher naturalness scores were associated with less trait distancing, *b* = -.04, *SE* = .02, *t*(153) = -1.45, *p* = .15. Apart from those reported above, no other effects of feedback condition, naturalness, or their interaction approached significance for trait ratings of either lesbians or gay men, all |*t*|'s < .7 all *p*'s > .5.

Exploratory Analysis

We ran exploratory analyses to examine the effect of participant gender. The only significant effects of gender involved ratings of gay men. Women (*M* = -.08, *SD* = .91), rated gay men more warmly compared to their ratings of other social groups than men (*M* = -.72, *SD* = 1.27, *t*(152) = 3.58, *p* < .001, *d* = .58), although both still showed an overall prejudice effect. There was also a significant main effect of gender for trait ratings of gay men in that men (*M* = 1.04, *SD* = .27), showed increased self-gay differences in trait ratings compared to women, *M* = .85, *SD* = .22, *t*(152) = -4.71, *p* < .001, *d* = .77. There were no significant interactions between gender and feedback condition on any measure of prejudice, all *F*'s < 1.05, all *p*'s > .3.

In sum, and consistent with Study 1, we found that participants who received feedback implying recategorization into a stigmatized outgroup showed decreased implicit prejudice. However, as predicted, this effect was moderated by naturalness beliefs such that only those who believed sexual orientation was biologically-determined and immutable showed this decrease. Similar to Study 1, there were no differences in explicit prejudice between the conditions, although we found tentative evidence for explicit distancing as a function of categorization threat for trait ratings of gay men.

General Discussion

We explored how learning of an immutable similarity implying membership in a stigmatized outgroup affected implicitly- and explicitly-assessed prejudice toward that group. We predicted (a) those who were given feedback implying a biological similarity indicating outgroup membership would exhibit reduced implicit prejudice (Studies 1 and 2), and (b) this effect would be moderated by participants' beliefs about the biological basis and immutability of group membership (Study 2). In both studies we found when dominant-group members (White participants in Study 1; straight participants in Study 2) were informed that they possessed biological similarities with a stigmatized outgroup (Black people in Study 1; "homosexuals" in Study 2), they showed reduced implicit prejudice. In Study 2, this effect was moderated by naturalness beliefs in that implicit prejudice reduction as a function of categorization threat occurred only among those who subscribed to essentialist (i.e., naturalness) beliefs about the identity.

Changes on Implicit but not Explicit Measures

While we found changes in implicitly-assessed prejudice as a function of categorization threat in accordance with predictions, we found only weak evidence of increased explicit

prejudice. Such implicit-explicit dissociations are not surprising (Greenwald, Poehlman, Uhlmann, & Banaji, 2009), and dual-process theories provide explanations for when and how such dissociations occur, as well as how they related to explicit judgments and behavior. Applied to the present findings, and consistent with other work (e.g., Olson & Fazio, 2006), our participants might have chosen to ignore their "gut feelings" when reporting their prejudices explicitly, and perhaps chose to report what they believed to be normatively appropriate responses. However, we did observe some evidence of explicit distancing from gay men in Study 2. Although we interpret this marginal effect with caution, the pattern of explicit prejudice across the present studies is consistent with evidence that contemporary American norms against overt expressions of racial prejudice are stronger than norms against the expression of prejudice against gay men (Crandall, Eshleman, & O'Brien, 2002; Zitek & Hebl, 2007). Thus, to the extent that categorization threat prompted explicit distancing, social norms may have, to some extent, dampened its overt expression.

The Role of Essentialist Beliefs

In Study 2, the effect of a categorization threat on implicit prejudice was moderated by naturalness beliefs, a subset of essentialist beliefs. Specifically, only those who believed sexual orientation to be biologically based and immutable showed reductions in implicit prejudice as a function of the threat. Such essentialist beliefs are related to "biomedicalization," the theory that many elements of social life are now understood in biomedical terms and from the perspective of advanced biotechnologies (cf. Kvaale, Haslam, & Gottdiener, 2013's use of the term "medicalization"). The sociological literature on biomedicalization has focused not only on how biomedicalization affects individuals' view of others, but how biomedicalization may influence our understanding of the self and social categories (Richardson, 2013; Waidzunas, 2015).

The present work demonstrates that while biomedical or essentialist explanations may potentially reduce prejudice toward stigmatized social groups, they also can have negative consequences. While genetic explanations have been found to reduce blame for individuals with mental illness, such beliefs are also associated with increases in perceptions of dangerousness and social distancing from these same individuals (Kvaale et al., 2013). Similarly, belief in the naturalness of sexual orientation categories is associated with less homonegativity, while beliefs in discreteness (i.e., social categories are distinct and non-overlapping), homogeneity (i.e., group members are similar to one another), and informativeness (i.e., group membership reveals fundamental things about the group members) are associated with more homonegativity (Grzanka et al., 2016). Beliefs about the naturalness of social categories implies reduced control by the individual for their group membership, which in turn is associated with less blame and more tolerance of those groups (Kvaale et al., 2013), while beliefs in the discreteness, homogeneity, and informativeness of social categories, on the other hand, suggests fundamental differences between groups and is associated with more prejudice (Grzanka et al., 2016).

Finally, it is important to note that while some essentialist beliefs may promote more 'tolerance' for different social groups, this 'tolerance' may not result in more 'acceptance.' Walters (2015) argued that biogenetic explanations of homosexual desires and behaviors may promote "tolerance" of sexual minorities more so than other forms of essentialism, but biogenetic explanations may not be a harbinger of widespread acceptance of non-normative sexualities or an appreciation for the social construction of sexual identities and desires. Furthermore, the resurgence of biomedical explanations of race and gender differences has been met with extensive controversy for the way that biogenetic essentialism may be used to justify

differential (i.e., unfair) treatment of women and people of color (Richardson, 2013). More research is needed to explore this possibility.

Illusory Ownership and Increased Associations with the Self

It is possible that individuals exhibit reduced implicit prejudice towards the stigmatized outgroup due to increased identification with that group (Greenwald et al., 2002). This recategorization may create an association between the self and the former outgroup, resulting in a positive association with that group (Cadinu & Rothbart, 1996). Previous research on interventions targeting individual's self-associations with outgroups has found reduced implicit prejudice and increased implicit self-associations. For example, in a study by Phills, Kawakami, Krusemark, and Nguyen (2017, Study 3), participants were assigned to either an intervention targeting racial attitudes or self-associations. Those in the racial attitudes intervention repeatedly associated Black people with positive evaluative concepts through evaluative conditioning. Those in the self-association intervention repeatedly associated the self with Blacks. Only participants in the self-association intervention showed both reduced implicit prejudice and increased implicit self-Black associations compared to other groups.

Research on illusory ownership of bodies has found that perceived ownership of an outgroup body or body part can result in decreased implicit prejudice towards that group. In a study by Peck, Seinfeld, Aglioti, and Slater (2013), participants saw either a virtual body or their own body that moved synchronously or asynchronously. While all participants who embodied virtual bodies expressed the same perceptions of body ownership, participants who embodied the synchronous dark skin body showed significantly reduced implicit prejudice (see also Farmer, Maister, & Tsakiris, 2013).

In all of these studies, participants showed more positive implicit associations when they associated the group with the self, whether through interventions targeting self-associations or perceived ownership of outgroup bodies. Applied to the present findings, it may be that our manipulation created a similar association such that receiving feedback implying identity recategorization led our participants to form an association between themselves with the outgroup which may have led to a decrease in implicit prejudice.

Preserving Self-Esteem?

Contemporary updates to balance theory assert that people are motivated to maintain consistency between their implicit attitudes, identities, and self-esteem (Heider, 1958; Greenwald et al., 2002). It may be that perceiving the stigmatized outgroup as more positive at the implicit level may have allowed our participants to maintain such a balance between their attitudes, self esteem, and identity.

Individuals associate their self-concept with the groups to which they belong, and since individuals tend to have positive associations with the self, they tend to associate their ingroups with positivity. According to this theory, people resist forming associations with concepts or valences that oppose associations they currently possess (Greenwald et al., 2002). For example, if people evaluate an outgroup as negative but evaluate themselves as positive, they are resistant to form an association between that outgroup and the self because this would create inconsistency or imbalanced triad in their attitudes.

In the current work, we forced dominant group members to associate their self-concepts with a stigmatized outgroup, which may have created an inconsistency between their implicit self and group attitudes. If people believe that group membership is immutable, they are unlikely to believe they can change the association between the self and outgroup. In order to maintain

consistency, individuals must resolve this discrepancy by either altering the valence associated with the self or with the outgroup. Since we can assume that individuals are unlikely to reduce their self-esteem by changing their self-evaluations from positive to negative, the only way to maintain affective-cognitive consistency is through a change in implicit outgroup attitudes—in this case, attitudes toward the minority group.

If self-esteem were central to the effect, we predict that those with the highest self-esteem would show the greatest prejudice reduction as a function of the categorization threat. In Study 2, we measured initial self-esteem and did find a significant 3-way interaction between condition, naturalness beliefs and explicit self-esteem, $b = -.56$, $SE = .16$, $t(148) = -3.60$, $p < .001$, $B = -12.37$, but the pattern of the interaction was uninterpretable. Furthermore, since we did not employ measures of implicit self-esteem or identity, we were unable assess any implicit changes in self-esteem or identity. Future research is needed to explore these possibilities.

Implications

Our findings clarify situations in which perceived similarity with outgroups reduces prejudice. Specifically, prejudice reduction might be facilitated when similarity is thought to be biological or immutable, or, relatedly, if the similarity implies change in one's social identities. Each of these entails a role of essentialist beliefs. By suggesting that essentialist beliefs may be relevant and useful for informing the reduction of prejudice across social identities, our work adds to the growing body of literature on intergroup relations and psychological essentialism (e.g., Mandalaywala, Amodio, & Rhodes, 2017). Indeed, another meaningful (and possibly more parsimonious) prejudice-reduction intervention could be to target individuals' essentialist beliefs in combination with other effective reduction strategies (i.e., intergroup dialogue). However, essentialist beliefs have inconsistent effects across social identity groups (Grzanka et al., 2016),

and future research should continue to uncover the conditions under which essentialist beliefs impact prosocial outcomes.

This research also has implications for contact-related interventions targeting self- and other-categorizations as a means of reducing prejudice. For example, the Common Ingroup Identity Model (CII; Gaertner & Dovidio, 2005) proposes that prompting members of disparate groups to consider membership in a common ingroup (e.g., "instead of Blacks and Whites, we are all Americans") can reduce prejudice. Tellingly, most research in this tradition has considered relatively non-essentialized (and often artificial) common ingroups. At least explicitly, CII manipulations depend on people's willingness to accept changes in self-categorization, including consideration of the plausibility of their membership in the new group. Thus, cognitively, there is some deliberation that impacts one's acceptance of a new common ingroup identity. If people's self-categorization beliefs are at odds with the new group membership, they will likely reject it. Such bounded effects of thoughtful consideration if identity-implicating novel information is consistent with Ferguson and colleagues' research (e.g., Mann & Ferguson, 2015) indicating that rapid changes in automatic evaluations occur only when novel information causes one to reconsider previous information and one has the time and cognitive resources to do so. We only observed implicit prejudice reduction among those who denied the malleability of these group memberships, leaving them "stuck" with a nonpreferred identity. Perhaps high naturalness beliefs promote greater consideration of identity threats and enhanced reconciliation processes between preferred identities and information suggesting possible membership in an outgroup, resulting in a stronger self-outgroup association.

In closing, the present work advances a potential moderating role for essentialist beliefs about pre-existing and newly-formed identities in the context of prejudice reduction

interventions involving self-recategorization. Moreover, our work suggests that further investigation of the complex interplay of multifarious essentialist beliefs and prejudice across identity groups is warranted.

References

Adams, H.E., Wright, L.W., & Lohr, B.A. (1996). Is homophobia associated with homosexual arousal? *Journal of Abnormal Psychology, 105,* 440-445. DOI:10.1037/0021-843X.105.3.440

Ashburn-Nardo, L., Voils, C. I., & Monteith, M.J. (2001). Implicit associations as the seeds of intergroup bias: how easily do they take root? *Journal of Personality and Social Psychology, 81,* 789-799. DOI:10.1037//0022-3514.81.5.789

Arseneau, J.R., Grzanka, P.R., Miles, J.R., & Fassinger, R.E. (2013). Development and initial validation of the sexual orientation beliefs scale (SOBS). *Journal of Counseling Psychology, 60,* 407-420. DOI:10.1037/a0032799

Beggan, J.K. (1992). On the social nature of nonsocial perception: The mere ownership effect. *Journal of Personality and Social Psychology, 62,* 229-237. DOI:10.1037/0022-3514.62.2.229

Bem, S.L. (1977). On the utility of alternative procedures for assessing psychological androgyny. *Journal of Consulting and Clinical Psychology, 45,* 196-205. DOI: 10.1037//0022-006X.45.2.196

Branscombe, N.R., Ellemers, N., Spears, R., & Doosje, B. (1999). The context and content of social identity threat. In N. Ellemers, R. Spears, & B. Doosje (Eds.), *Social identity: Context, commitment, content* (pp. 35-58). Malden, MA: Blackwell.

Brehm, J.W. (1956). Post-decision changes in the desirability of alternatives. *The Journal of Abnormal Psychology, 52,* 384-389. DOI: 10.1037/h0041006

Cadinu, M.R. & Rothbart, M. (1996). Self-anchoring and differentiation processes in the minimal group setting. *Journal of Personality and Social Psychology, 70,* 661-677. DOI: 10.1037/0022-3514.70.4.661

Cialdini, R.B., Borden, J.B., Thorne, A., Walker, M.R., Freeman, S., & Sloan, L.R. (1976). Basking in reflected glory: Three (football) field studies. *Journal of Personality and Social Psychology, 34,* 366-375. DOI: 10.1037//0022-3514.34.3.366

Cone, J., & Ferguson, M. J. (2015). He did what? The role of diagnosticity in revising implicit evaluations. *Journal of Personality and Social Psychology, 108,* 37–57. DOI: 10.1037/pspa0000014

Crandall, C.S., Eshleman, A., & O'Brien, L. (2002). Social norms and the expression and suppression of prejudice: The struggle for internalization. *Journal of Personality and Social Psychology, 82,* 359-378. DOI: 10.1037//0022-3514.82.3.359

Ellemers, N., Wilke, H., & van Knippenberg, A. (1993). Effects of the legitimacy of low group or individual status on individual and collective status-enhancement strategies. *Journal of Personality and Social Psychology, 64,* 766-778. DOI: 10.1037/0022-3514.64.5.766

Farmer, H., Maister, L., & Tsakiris, M. (2013). Change my body, change my mind: the effects of illusory ownership of an outgroup hand on implicit attitudes toward that outgroup. *Frontiers in Psychology, 4,* 1016. DOI: 10.3389/fpsyg.2013.01016

Fazio, R.H., Jackson, J.R., Dunton, B.C., & Williams, C.J. (1995). Variability in automatic activation as an unobtrusive measure of racial attitudes: A bona fide pipeline? *Journal of Personality and Social Psychology, 69,* 1013–1027.

Fazio, R.H., & Olson, M.A. (2014). The MODE model: Attitude-Behavior Processes as a Function of Motivation and Opportunity. In Sherman, J.W., Gawronski, B., & Trope, Y. (Eds.). *Dual process theories of the social mind*. New York: Guilford Press.

Gaertner, S. L., & Dovidio, J. F. (2005). Understanding and addressing contemporary racism: From aversive racism to the common ingroup identity model. *Journal of Social issues, 61,* 615-639. DOI:10.1111/j.1540-4560.2005.00424.x

Gawronski, B., Bodenhausen, G.V., & Becker, A.P. (2007). I like it, because I like myself: Associative self-anchoring and post-decisional change of implicit evaluations. *Journal of Experimental Social Psychology, 43,* 221-232. DOI:10.1016/j.jesp.2006.04.001

Gawronski, B., & Bodenhausen, G.V. (2006). Associative and propositional; processes in evaluation: An integrative review of implicit and explicit attitude change. *Psychological Bulletin, 132,* 692-731. DOI:10.1037/0033-2909.132.5.692

Gibbons, F.X. (1985). Stigma perception: Social comparison among mentally retarded persons. *American Journal of Mental Deficiency, 90,* 98-106.

Greenwald, A.G., Banaji, M.R., Rudman, L.A., Farnham, S.D., Nosek, B.A., & Mellott, D.S. (2002). A unified theory of implicit attitudes, stereotypes, self-esteem, and self-concept. *Psychological Review, 109,* 3-25. DOI: 10.1037//0033-295X.109.1.3

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. DOI:10.1037/0022-3514.85.2.197

Greenwald, A.G., Poehlman, T.A., Uhlmann, E.L., & Banaji, M.R. (2009). Understanding and using the Implicit Association Test: III. Meta-analysis of predictive validity. *Journal of Personality and Social Psychology, 97,* 17-41. DOI:10.1037/a0015575

Grzanka, P.R., Zeiders, K.H., & Miles, J.R. (2016). Beyond "born this way?" Reconsidering sexual orientation beliefs and attitudes. *Journal of Counseling Psychology, 63,* 67-75. DOI: 10.1037/cou0000124

Haslam, N., Rothschild, L., & Ernst, D. (2000). Essentialist beliefs about social categories. *British Journal of Social Psychology, 39,* 113–127. DOI:10.1348/014466600164363

Haslam, N., Rothschild, L., & Ernst, D. (2002). Are essentialist beliefs associated with prejudice?. *British Journal of Social Psychology, 41,* 87-100. DOI:10.1348/014466602165072

Haslam, N., & Levy, S.R. (2006). Essentialist beliefs about homosexuality: Structure and implications for prejudice. *Personality and Social Psychology Bulletin, 32,* 471-485. DOI:10.1177/0146167205276516

Hegarty, P., & Pratto, F. (2001). Sexual orientation beliefs: Their relationship to anti-gay attitudes and biological determinist arguments. *Journal of Homosexuality, 41,* 121-135. DOI:10.1300/J082v41n01_04

Heider, Fritz (1958). *The Psychology of Interpersonal Relations*. New York, NY: Wiley.

Hofmann, W., Gawronski, B., Gschwendner, T., Lê, H.H., & Schmitt, M. (2005). A meta-analysis on the correlation between the implicit association test and explicit self-report measures. *Personality and Social Psychology Bulletin, 31,* 1369-1385. DOI:10.1177/0146167205275613

Jarvis, B.G. (2014). MediaLab (Version 2014) [Computer Software]. New York, NY: Empirisoft Corporation.

Jarvis, B.G. (2014). DirectRT (Version 2014) [Computer Software]. New York, NY: Empirisoft Corporation.

Kendrick, R.V., & Olson, M.A. (2012). When feeling right leads to being right in the reporting of implicitly-formed attitudes, or how I learned to stop worrying and trust my gut. *Journal of Experimental Social Psychology, 48,* 1316-1321. DOI:10.1016/j.jesp.2012.05.008

Koole, S.L., Dijksterhuis, A., & van Knippenberg, A. (2001). What's in a name: Implicit self-esteem and the automatic self. *Journal of Personality and Social Psychology, 80,* 669-685. DOI:10.1037/0022-3514.80.4.669

Kvaale, E.P., Haslam, N., & Gottdiener, W.H. (2013). The 'side effects' of medicalization: A meta-analytic review of how biogenetic explanations affect stigma. *Clinical Psychology Review, 33,* 782-794. DOI:10.1016/j.cpr.2013.06.002

Long, K., & Spears, R. (1997). The self-esteem hypothesis revisited: Differentiation and the disaffected. In R. Spears, P.J. Oakes, N. Ellemers, & S.A. Haslam (Eds.), *The social psychology of stereotyping and group life* (pp. 296-317). Oxford, UK: Blackwell.

Mandalaywala, T., Amodio, D.M., & Rhodes, M. (2017). Essentialism promotes racial prejudice by increasing endorsement of social hierarchies. *Social Psychological and Personality Science, 9,* 461-469. DOI:10.1177/1948550617707020

Mann, T. C., & Ferguson, M. J. (2015). Can we undo our first impressions? The role of reinterpretation in reversing implicit evaluations. *Journal of Personality and Social Psychology, 108,* 823–849. DOI: 10.1037/pspa0000021

Medin, D.L. (1989). Concepts and conceptual structure. *American Psychologist, 44,* 1469-1481. DOI:10.1037/0003-066X.44.12.1469

Meindl, J.R., & Lerner, M.J. (1984). Exacerbation of extreme responses to an out-group. *Journal of Personality and Social Psychology, 47,* 71-84. DOI:10.1300/J082v43n02_02

Morrison, M.A., & Morrison, T.G. (2003). Development and validation of a scale measuring modern prejudice toward gay men and lesbian women. *Journal of Homosexuality, 43,* 15-37. DOI:10.1300/J082v43n02_02

Nosek, B.A., Banaji, M.R., & Greenwald, A.G. (2002). Harvesting implicit group attitudes and beliefs from a demonstration website. *Group Dynamics, 6,* 101-115. DOI:10.1037//1089-2699.6.1.101

Novak, D.W., & Lerner, M.J. (1968). Rejection as a consequence of perceived similarity. *Journal of Personality and Social Psychology, 9,* 147-152. DOI:10.1037/h0025850

Nuttin, J.M., Jr. (1985). Narcissism beyond Gestalt and awareness: The name letter effect. *European Journal of Social Psychology, 15,* 353-361. DOI:10.1002/ejsp.2420150309

Olson, M.A., & Fazio, R.H. (2001). Implicit attitude formation through classical conditioning. *Psychological Science, 12,* 413–417. DOI:10.1111/1467-9280.00376

Olson, M.A., & Fazio, R.H. (2004). Reducing the Influence of Extrapersonal Associations on the Implicit Association Test: Personalizing the IAT. *Journal of Personality and Social Psychology, 86,* 653–667. DOI:10.1037/0022-3514.86.5.653

Olson, M.A., & Fazio, R.H. (2006). Reducing automatically-activated racial prejudice through implicit evaluative conditioning. *Personality and Social Psychology Bulletin, 32, 421-433.* DOI: 10.1177/0146167205284004

Olson, M.A., & Zabel, K.L. (2016). Measures of prejudice. In T. Nelson (Ed.), *Handbook of Prejudice, Stereotyping, and Discrimination* (pp. 175-212). New York: Psychology Press.

Otten, S., & Wentura, D. (1999). About the impact of automaticity in the Minimal Group Paradigm: evidence from affective priming tasks. *European Journal of Social*

Psychology, 29, 1049-1071. DOI:10.1002/(SICI)1099-0992(199912)29:8<1049::AID-EJSP985>3.0.CO;2-Q

Peck, T.C., Seinfeld, S., Aglioti, S.M., & Slater M. (2013). Putting yourself in the skin of a black avatar reduces implicit racial bias. *Conscious Cognition, 22*, 779-787. DOI:10.1016/j.concog.2013.04.016.

Panofsky, A., & Donovan, J. (2017). When Genetics Challenges a Racist's Identity: Genetic Ancestry Testing among White Nationalists. *SocArXiv*, Retrieved from: https://osf.io/preprints/socarxiv/7f9bc/

Perdue, C.W., Dovidio, J.F., Gurtman, M.B., & Tyler, R.B. (1990). Us and them: Social categorization and the process of intergroup bias. *Journal of Personality and Social Psychology, 59*, 475-486. DOI:10.1037/0022-3514.59.3.475

Phills, C.E., Kawakami, K., Krusemark, D.R., & Nyguen, J. (2017). Does Reducing Implicit Prejudice Increase Out-Group Identification? The Downstream Consequences of Evaluative Training on Associations Between the Self and Racial Categories. *Social Psychological and Personality Science.* DOI:10.1177/1948550617732817

Richardson S. S. (2013). *Sex itself: The search for male and female in the human genome*. Chicago, IL: University of Chicago Press.

Rosenberg, M. (1979). Conceiving the Self. New York: Basic Books.

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*(5), 867-878. DOI: 10.1002/ejsp.393

Schimel, J., Pyszczynski, T., Greenberg, J., O'Mahen, H., & Arndt, J. (2000). Running from the shadow: Psychological distancing from others to deny characteristics people fear in themselves. *Journal of Personality and Social Psychology, 78*, 446-462. DOI:10.1037//0022-3514.78.3.446

Tajfel, H., Billig, M.G., Bundy, R.P. & Flament, C. (1971). Social categorization and intergroup behaviour. *European Journal of Social Psychology*, *1*, 149-177. DOI:10.1002/ejsp.2420010202

Tajfel, H., & Turner, J.C. (1979). An integrative theory of inter-group conflict. In W.G. Austin & S. Worchel (Eds.), The social psychology of inter-group relations (33–47). Monterey, CA:Brooks/Cole.

Turner, J.C. (1999). Some current issues in research on Social Identity and Self-categorization Theories. In N. Ellemers, R. Spears & B. Doosje (Eds.), *Social identity: Context, commitment, content* (pp. 6-34). Malden, MA: Blackwell.

Waidzunas, T. (2015). *The straight line: How the fringe science of ex-gay therapy reoriented sexuality*. Minneapolis, MN: University of Minnesota Press.

Walther, E., & Trasselli, C. (2003). I like her, because I like myself: Self-evaluation as a source of interpersonal attitudes. *Experimental Psychology, 50,* 239-246. DOI:10.1026//1618-3169.50.4.239

Walters, S.D. (2015). *The tolerance trap: How god, genes, and good intentions are sabotaging gay equality*. New York, NY: New York University Press.

Wolsko, C., Park, B., Judd, C.M., & Wittenbrink, B. (2000). Framing interethnic ideology: Effects of multicultural and color-blind perspectives on judgments of groups and individuals. *Journal of Personality and Social Psychology, 78,* 635-654. DOI:10.1037/0022-3514.78.4.635

Yzerbyt, V.Y., Rocher, S., & Schadron, G. (1997). Stereotypes as explanations: A subjective essentialistic view of group perception. In R. Spears, P. Oakes, N. Ellemers, & A. Haslam (Eds.), *The psychology of stereotyping and group life* (pp. 20-50). London: Basil Blackwell.

Zárate, M.A., Garcia, B., Garza, A.A., & Hitlan, R.T. (2004). Cultural threat and perceived realistic group conflict as dual predictors of prejudice. *Journal of Experimental Social Psychology, 40,* 99-105. DOI:10.1016/S0022-1031(03)00067-2

Zitek, E.M., & Hebl, M.R. (2007). The role of social norm clarity in the influenced expression of prejudice over time. *Journal of Experimental Social Psychology, 43,* 867-876. DOI:10.1016/j.jesp.2006.10.010

Study 1

Table 1. Bivariate Correlations between Dependent Measures

Table 1
Bivariate Correlations between Measures

Parameter	1	2	3
IAT	1.0		
Feeling Thermometer	-.26*	1.0	
Trait Distancing	.07	-.18	1.0

Note: ** indicates significant at the .01 level (2-tailed); * indicates significant at the .05 level (2-tailed).

Study 2

Table 2. Bivariate Correlations between Dependent Measures in Study 2

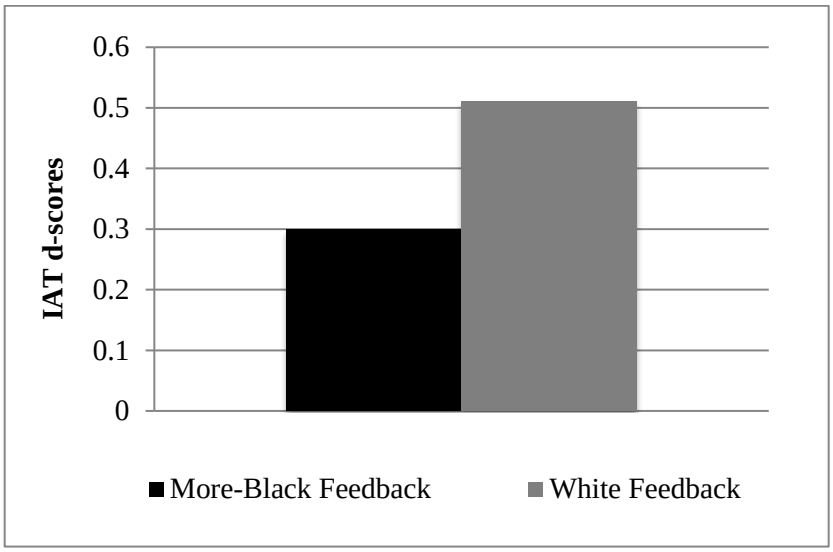
Table 2
Bivariate Correlations between Dependent Measures

Parameter	1	2	3	4	5
IAT	1.0				
Feeling Thermometer: Gay Men	-.36**	1.0			
FT: Lesbian Women	-.28**	.66**	1.0		
Trait Distancing: Gay Men	.18*	-.33*	-.11	1.0	
Trait Distancing: Lesbian Women	.34**	-.36**	-.33**	.54**	1.0

Note: ** indicates significant at the .01 level (2-tailed); * indicates significant at the .05 level (2-tailed).

Study 1

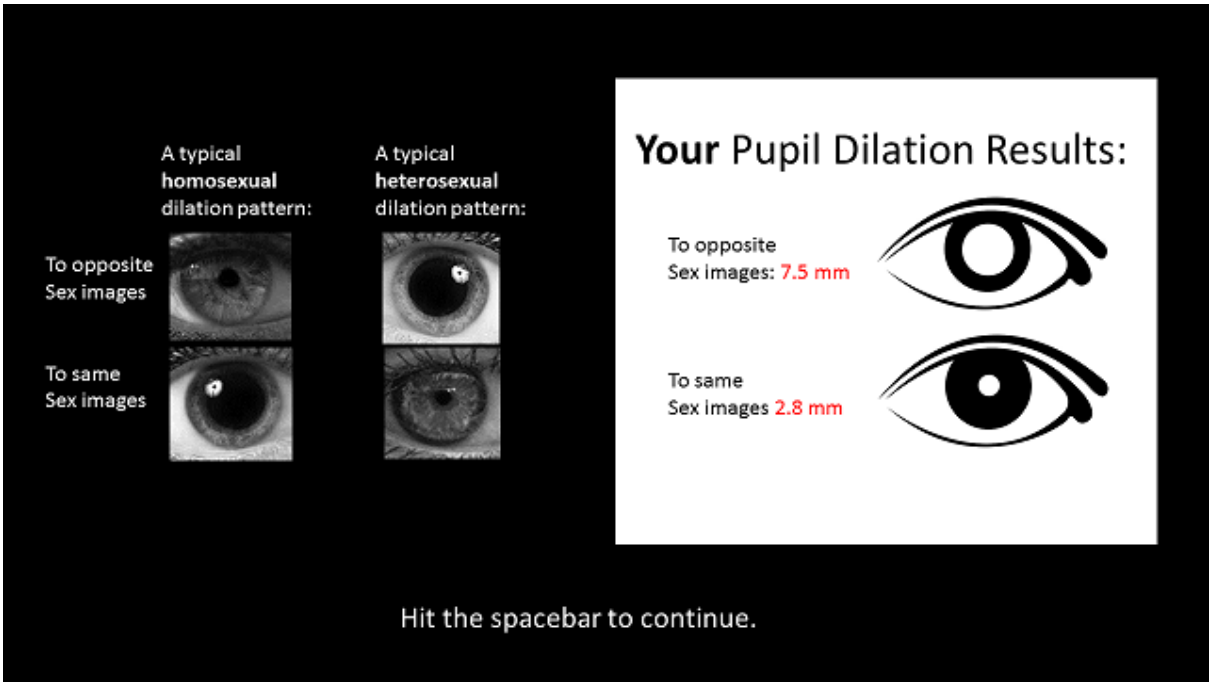
Figure 1. Effect of Feedback Condition on IAT d-scores



Note: Positive numbers indicate an implicit preference for Whites over Blacks.

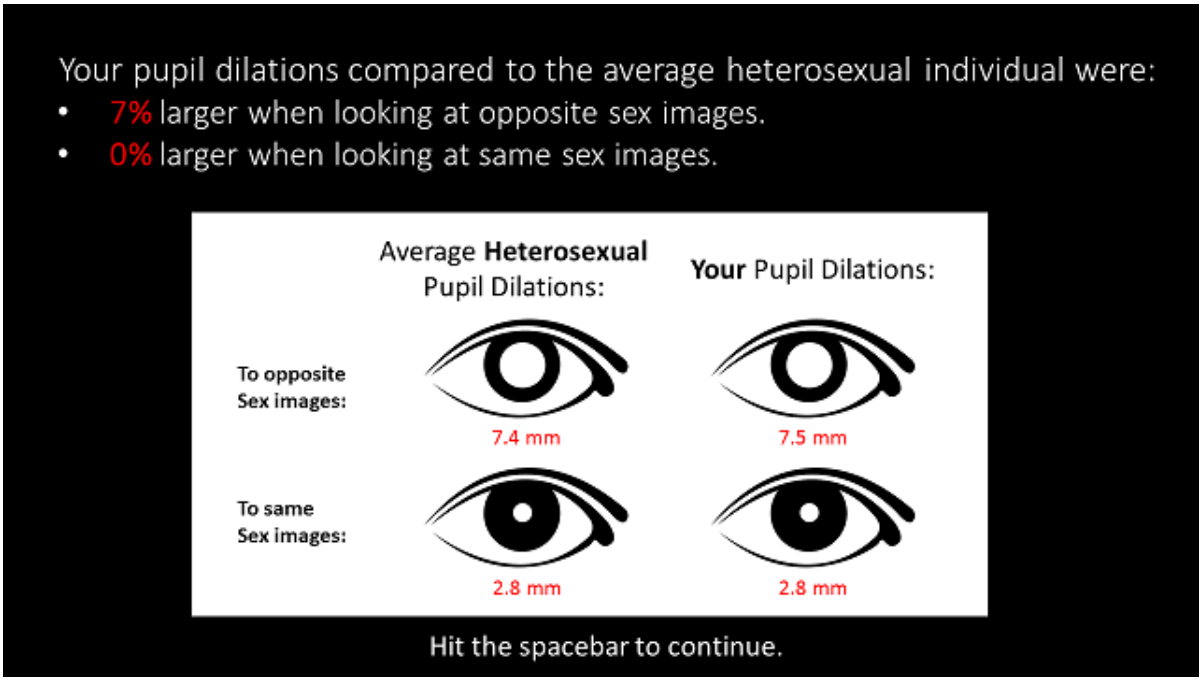
Study 2

Figure 2.1. Straight Condition Pupil Dilation Feedback 1



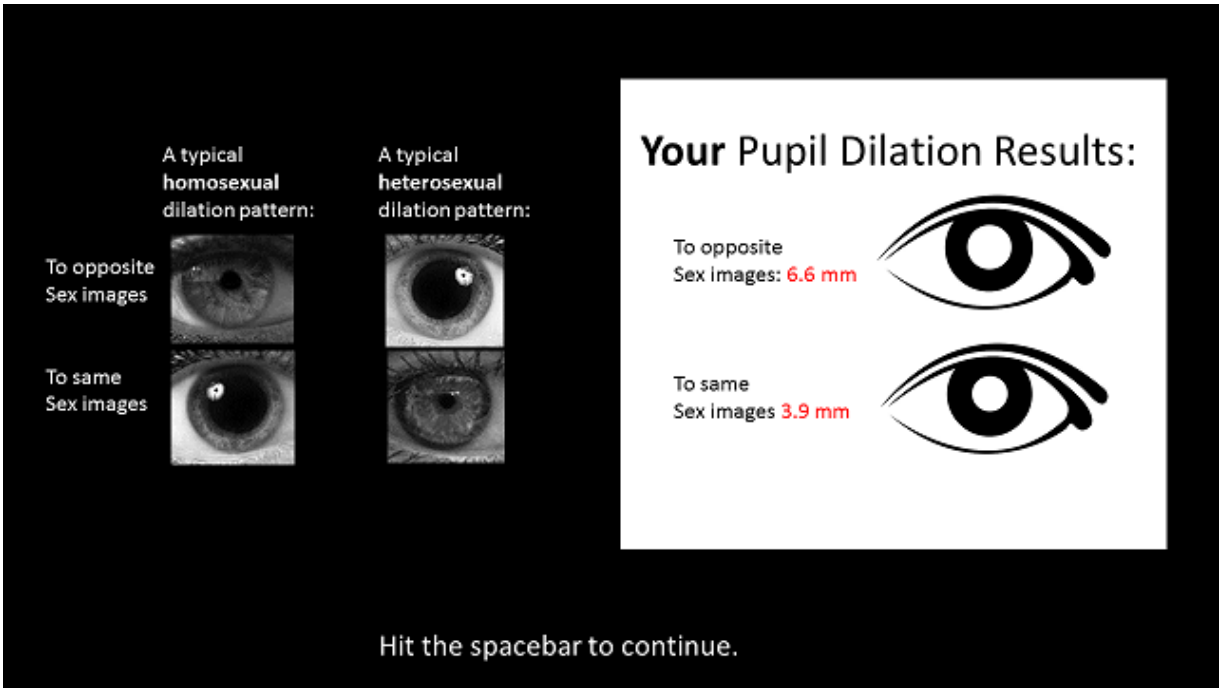
Study 2

Figure 2.2. Straight Condition Pupil Dilation Feedback 2



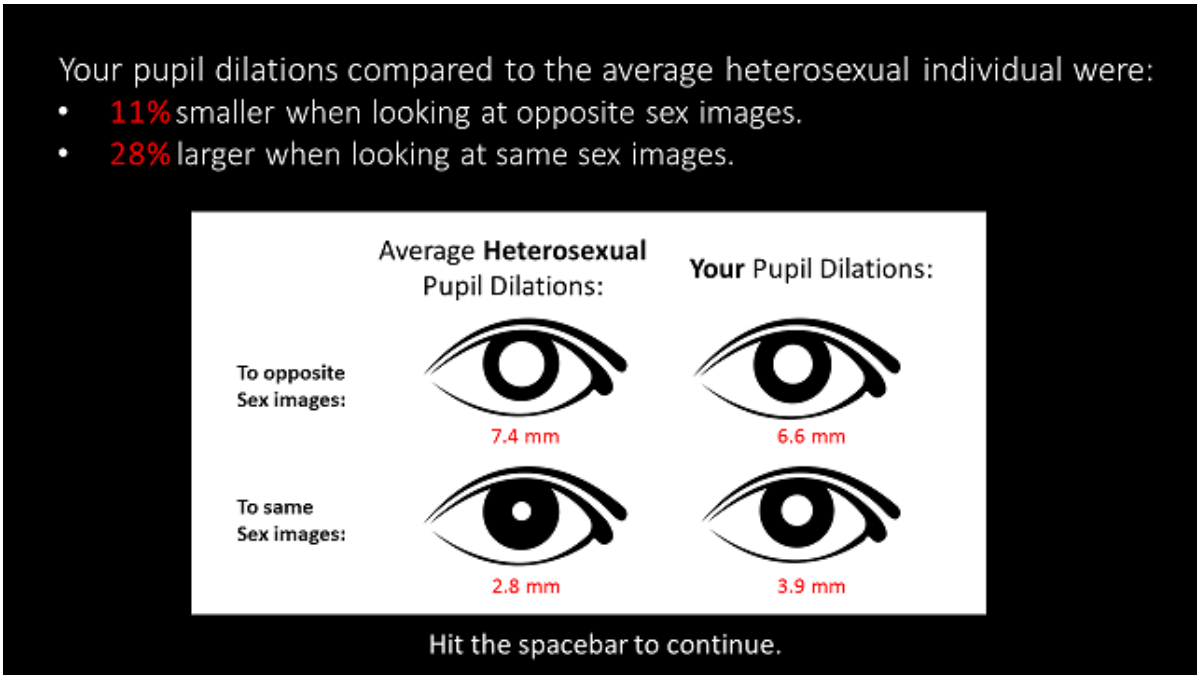
Study 2

Figure 2.3. More-Gay Condition Pupil Dilation Feedback 1



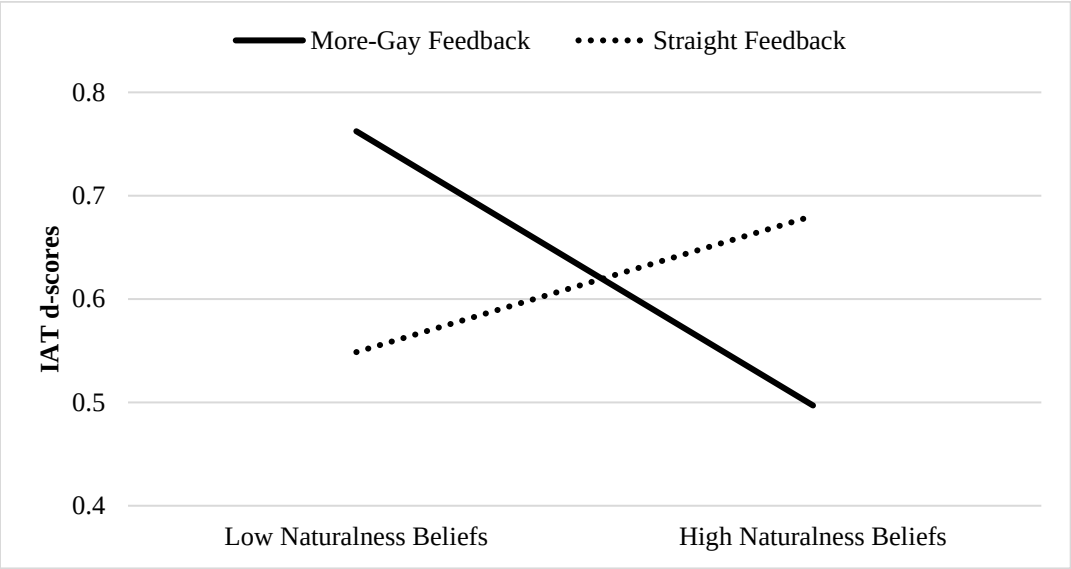
Study 2

Figure 2.4. More-Gay Condition Pupil Dilation Feedback 2



Study 2

Figure 3. Effects of Feedback Condition and Naturalness Beliefs on IAT d-scores



Note: Positive numbers indicate an implicit preference for straight over gay and lesbian individuals.