International Journal of Psychological Studies; Vol. 12, No. 4; 2020
ISSN 1918-7211 E-ISSN 1918-722X
Published by Canadian Center of Science and Education

Race is Still Black and White: Voluntary Racial Phenotypic Change Elicits Meaning Threat and Backlash

Jordan Seliger[1] & Avi Ben-Zeev[2]

[1] Department of Neurology and Neurological Sciences, Stanford University, USA

[2] Department of Psychology, San Francisco State University, USA

Correspondence: Jordan Seliger, Department of Neurology and Neurological Sciences, Stanford University, 300 Pasteur Drive, Palo Alto, CA 94304, USA.

Received: September 10, 2020 Accepted: October 20, 2020 Online Published: October 29, 2020

doi:10.5539/ijps.v12n4p1 URL: https://doi.org/10.5539/ijps.v12n4p1

Abstract

We offer evidence that a target who voluntarily changes his/her racial phenotypic features causes perceivers to engage in two-pronged social policing of racial group boundaries: (a) vilifying and disliking the target (cognitive and affective backlash; external policing) (Experiments 1a-1b, 2, & 3) and (b) increasing own racial essentialism, in response to a meaning threat (internal policing) (Experiment 3). In all experiments, participants received a vignette of a protagonist that underwent non-elective surgery (white/Asian, Experiments 1a-1b; white/Black, Experiments 2-3). In the *voluntary change* condition, the protagonist asks that the surgeon change his/her racial features to resemble that of a different race whereas, in the *involuntary change* condition the protagonist asks that the surgeon keep his/her racial features intact (Experiment 1: eye shape, Experiment 2: Afrocentric features). Findings supported the predictions and showed a dissociation between similarity and categorization judgments, underscoring the essentialized versus socially constructed nature of beliefs about race.

Keywords: race, essentialism, dehumanization, social categorization

1. Introduction

People who undergo ethnic cosmetic procedures opt to change their racial phenotypic features. In Asian, Latino/a and Black populations, these procedures often include creasing eyelids, sharpening noses, thinning lips, and bleaching skin to achieve a more European/Anglo appearance (Davis, 2003; Hunter, 2005; Hunter, 2011). Given a social reality in which racial phenotypic features are perceived as "diagnostic" of hidden psychological properties - such that non-white (versus white) phenotypes are often associated with unwholesome and negative characteristics (for *racial phenotypicality bias*, see Maddox, 2004) (for *colorism*, see Hunter, 2007) - it may not be surprising, therefore, that the majority of ethnic cosmetic procedures are geared towards and undertaken by people of color. Hunter (2011) argued that people of color who opt to change their racial phenotypes to appear whiter tend to maintain their racial identities as intact: "The racial capital of whiteness is now something consumers can buy. It is not necessarily the case that consumers of skin-whitening products want to be white per se, but the huge demand for these products suggests that many people want to look white, or at least light, relative to other people in their racial or ethnic group" (Hunter, 2011, p. 149).

Skin bleaching, in particular, has grown into a multibillion-dollar industry despite the fact that many products contain chemical compounds (e.g., mercury, hydroquinone, and corticosteroids) linked to adrenal gland damage, kidney and liver failure as well as cancer and have been therefore prohibited for sale (but not for manufacturing) in the United States (see, Hunter, 2011). Advertisements for these products emphasize notions of "white beauty," privilege a whiter over darker appearance (e.g., "dark out, white in"), and urge users to "Click LIKE for fairer skin" on social media, such as on Facebook (Doland, 2014). Although less frequently, white individuals have also been documented to voluntarily alter racial phenotypic features via procedures such as filling lips, augmenting buttocks, and darkening skin tone to look more "exotic" (see Dreisinger, 2008).

The present set of studies are centered on the following question: Do people who voluntarily change their racial phenotypic features to appear more similar to a different race's become subject to social backlash, that is, to negative attitudes and dislike in the ultimate service of maintaining essentialist status quo beliefs about race? (For backlash, see Rudman & Fairchild, 2004; also see Rudman & Phelan, 2008; Rudman, Moss-Racusin, Phelan &

Nauts, 2012.) At first blush, it seems reasonable to predict that in a culture that privileges a white appearance, any negative societal reaction would be directed at a person who voluntarily changes away from (versus changes towards) an Anglo/Eurocentric phenotype. We contend, however, that regardless of whether the direction of racial phenotypic change aligns or misaligns with a Eurocentric appearance, targets who voluntarily alter their racial phenotypic features would incur cognitive and affective backlash (herein, vilification and dislike, respectively). The underlying rationale is that such individuals challenge the "social and representational authority of 'race'" (Wald, 2000, p. 4), that is, the set of culturally constructed beliefs that link differences in phenotypic features to a biological/essentialist and hierarchical categorization (Smedley & Smedley, 2005).

Race is a highly essentialized category, such that members of a given group are perceived to share a hidden causal property, or 'essence,' which is thought to constrain group members' surface features, bind category members together, and explain their racial identity (e.g., Gelman, 2003; Haslam, Rothschild & Ernst, 2000; Prentice & Miller, 2007). This conceptualization entails that race is a biological category and that its boundaries are immutable (i.e., if one is born into the category, one stays a member of that category) (e.g., Haslam et al., 2000). Given the conceptualization of race as biological and fixed, we predict a person who opts to change skin-deep, but socially-laden, racial phenotypic features to become more similar to a different race's will cause perceivers to engage in social policing of racial group boundaries via vilification and dislike of the target, which constitute cognitive and affective backlash, respectively (for backlash, see Rudman & Fairchild, 2004; also see Rudman & Phelan, 2008; Rudman, Moss-Racusin, Phelan & Nauts, 2012).

Backlash entails punishing norm violators via social sanctions, while aiming to deter others, in an attempt to maintain existing beliefs and social hierarchies (Rudman, 1998; Rudman & Fairchild, 2004; Rudman & Phelan, 2008; Rudman, Moss-Racusin, Phelan, & Nauts, 2012). When perceivers encounter targets that violate social expectations (e.g., agentic women who defy gender-based stereotypes), perceivers tend to experience dislike towards these targets (e.g., Rudman et al., 2012). Dislike is often a manifestation of prejudice (Fiske, 2002) and is thus not a trivial social sanction. Being disliked causes people to evaluate themselves more poorly (Srivastava & Beer, 2005), experience discrimination in the workplace (Rudman & Glick, 2001), among other adverse outcomes. In the context of race, it follows that a target that opts to change his or her racial phenotypic features to appear more similar to a different race's and thus transgresses race, would be similarly met with dislike.

Moreover, we contend that a target that transgresses race will incur an additional backlash in the form of vilification. We situate the construct of vilification in theorizing and empirical findings on humanness and subtle and everyday forms of dehumanization (Haslam, 2006). In particular, Haslam and colleagues (Bastian & Haslam, 2010; Haslam, 2006) have shown that there exists a set of positively and negatively valenced (e.g., conscientiousness and stinginess) human uniqueness traits, which are conceived as separating humans from other animals (Bastian & Haslam, 2010) (also see Leyens et al., 2001 on primary and secondary emotions). Vilification is a judgment of 'core badness' and is thus conceived herein as a simultaneous exaggeration of a person's negative human uniqueness traits and underestimation of his/her positive human uniqueness traits (versus *infrahumanization* in which judgments of both positive and negative human uniqueness traits are reduced) (for perceptual dehumanization, in which norm violators' faces are processed less holistically, see Fincher & Tetlock, 2016).

The mere existence of a target that voluntarily changes their racial phenotypic features calls into question the immutability or fixity beliefs about race. Thus, punishing the transgressor via backlash ("you are bad and I don't like you") might not completely resolve the uncertainty about the state of the world that perceivers often experience when they encounter norm transgressions (i.e., *meaning threat*, see Heine, Proulx, and Vohs, 2006), and would therefore necessitate an additional internal resolution, such as in the form of strengthening status quo essentialist beliefs about race. A meaning threat can be illustrated metaphorically by Yoko Ono's classic art piece, entitled *White Chess Set* (1966), which consists of an all-white board with all-white pieces. Given that both players' pieces cannot be distinguished from each other's – even if one set of pieces were conceived of as black pieces that had been painted over in white – this game will inevitably lead to a confusion between 'us' and 'them' pieces, rendering the distinction of opposing groups and the game itself meaningless.

We thus hypothesize that beyond backlash, perceivers will react to a person who opts to change his/her racial phenotypic features by experiencing a meaning threat, and in particular, feelings of uncertainty about the fixity and predictability of 'race.' Social policing of a voluntary racial transgressor is aligned with people's documented desire to restore certitude in the status quo and a form of "fluid compensation" (see, Heine, Proulx & Vohs, 2006) - an attempt to resolve the need for cognitive closure (Kruglanski & Webster, 1996) (also see Jost & Banaji, 1994; Wakslak, Jost & Bauer, 2011 for a system justification perspective). The target's volition is central to these

predictions because accidental (versus voluntary) change is less likely to threaten status quo beliefs given that people who incur them still abide by social norms (i.e., are victims of circumstance).

2. Experimental Paradigm and Predictions

In all experiments, participants received a vignette of a protagonist that underwent non-elective surgery (white/Asian, Experiments 1a-1b; white/Black, Experiments 2-3). In the *voluntary change* condition, the protagonist asks that the surgeon change his/her racial features to resemble that of a different race whereas, in the *involuntary change* condition, the protagonist asks that the surgeon keep his/her racial features intact (Experiment 1: eye shape, Experiment 2: 'Afrocentric' features and skin tone). In the control condition, the protagonist did not incur any change. We predicted that a voluntary (versus accidental) change to the protagonist's racial phenotypic features would invoke social policing, that is, cognitive and affective backlash in the form of vilification and dislike.

Essentialism is predicated on a binary structure: the essence is assumed to be present or absent (Medin & Ortony, 1989). Thus, we predict that volition would have an effect on backlash but not on categorization judgments. That is, the 'transgressor' will still be considered to be a member of the racial category of origin, albeit, a more poorly functioning exemplar that is vilified and disliked. Thus, regardless of whether phenotypic change is voluntary or accidental, we expect to find a dissociation between similarity and categorization judgments, such that post-phenotypic change, the protagonist will be judged as more similar to the new racial category but as belonging to the original racial category. So far, this form of "origin essentialism" has only been shown with fictitious, non-human exemplars that did not possess volition (Rips, 1989).

Taken together, if predictions are corroborated, the findings would help shed light on an under-researched topic: the nature of societal reactions to individuals who voluntarily change their racial phenotypic features. In particular, the social pressure placed on people of color to appear whiter - and a multibillion-dollar industry that profits from it (in the face of documented health risks, see Hunter 2011) - might lead to a "damned if you do; damned if you don't" consequence for these individuals, concomitant with heightened cultural conceptions that reify race, serve to police racial boundaries, and maintain the status quo. Such findings would add to the extant literature on the entrenched nature of racial essentialism, a belief system which is at odds with a reality in which race is socially constructed (e.g., Bodmer & Cavalli-Sforza, 1976; Molnar, 1992; Gould, 1981; Tate & Audette, 2001) and that has been linked to stereotyping and social inequities (e.g., Bastian & Haslam, 2006; Prentice & Miller, 2007).

2.1 Experiment 1a

Experiment 1a was designed to explore whether a voluntary (versus accidental) change to the protagonist's racial phenotypic features would invoke cognitive and affective backlash in the form of vilification and dislike (external policing). European and Asian American participants received vignettes that depicted an Asian/white man/woman. In the voluntary and involuntary change conditions, the protagonist's eye shape was changed (on the racial diagnosticity of eye shape, see Brown, Dane, & Durham, 1998), whereas in the control condition, the protagonist incurred no change. We predicted that a voluntary (versus accidental) racial phenotypic change would invoke cognitive and affective backlash in the form of vilification and dislike. Our investigation was designed to explore the global nature of social policing to voluntary phenotypic transgressors given that racial essentialism is foundational to human categorization (Prentice & Miller, 2007). We deemed it important, however, to select a similar number of Asian- and white-identified participants to prevent the findings from being biased by an unequal representation of participant race.

Across all experiments, based on a prospective power analysis (Faul, Erdfelder, Lang, & Buchner, 2007), we determined that for a sample of 93 participants per condition ($\alpha = .05$), we could detect an effect size of $\eta^2 = .10$ with a probability of .80, in a 2 (protagonist's race of origin: Black/Asian versus white) x 3 (Condition: control, voluntary, and involuntary) between-subjects design. However, the current paradigm is novel and, therefore, a true effect size is unknown. Thus, we aimed to include 200 participants (and up to a maximum of 257 participants) per experiment. Data for all experiments can be found on the Open Science Framework (OSF): https://osf.io/wh6fr/?view_only=d2c5cf953b8a4784bb57c45faa2fafb0.

Method

Participants

Two hundred seventy-five participants were recruited via TurkPrime, a crowd-sourcing website that uses Amazon.com's MTurk. All participants had an MTurk approval rating of 90% or higher and lived in the United States. Participants received $.50 for their participation in the study. Twenty-nine participants were excluded (14 voluntary condition, 8 involuntary condition, 7 control condition) for either incorrectly answering manipulation

checks (described below), missing data, not identifying as Asian or white, or for not allowing their data to be used, leaving 246 participants (111 male, 135 female; 128 Asian-identified, 118 white-identified; M_{Age} = 34.9, SD_{Age} = 10.8) in the final analyses. Removing the 29 participants did not influence the results, so they were excluded from the final analysis.

Design

We employed a 2 (protagonist's race of origin: Asian versus white) x 3 (condition: control, voluntary, and involuntary) between-subjects factorial design.

Materials

Vignettes.

Control.

John/Jane is a health-conscious Asian American/White man/woman in his/her early 30s with a successful career, a bright outlook on life, and no family history of serious illness.

Involuntary versus voluntary change. The first part was identical to the control vignette. The second part was as follows:

One day, John/Jane contracted an eye infection and surgery was required to preserve his/her vision. He/she underwent surgery by an ophthalmologist who was also a plastic surgeon. The surgeon gave John/Jane the option of keeping his/her eyes the same or changing them.

Involuntary change. John/Jane requested that his/her eyes stay the same. The surgeon tried to comply with his/her request, but due to a surgical error his/her eyes ended up looking like a White/Asian person's.

Voluntary change. John/Jane requested that his/her eyes be changed to look like a White/Asian person's. The surgeon complied with his/her request and his/her eyes ended up looking like a White/Asian person's.

Categorization vs. similarity dissociation measure. Two items were adapted from Rips (1989) to measure categorization, "How would you categorize John/Jane's race?" and similarity, "What race would you physically describe John/Jane as being most similar to?" using four-point Likert-type scales (1 = Asian; 2 = More Asian than White; 3 = More White than Asian; 4 = White). Ratings were recoded such that lower values corresponded to the protagonist's race of origin. The order of items and scale points were counterbalanced across all participants.

Vilification. This measure consisted of ten items adapted from Bastian and Haslam (2010). Participants were asked to rate the extent to which Jane/John possessed the following traits: broadminded, conscientious, humble, polite, and thorough (five positive uniquely human traits; α = .864) as well as, disorganized, hard-hearted, ignorant, rude, and stingy (five negative uniquely human traits; α = .889) on seven-point Likert scales (1 = not at all; 7 = very much so).

Likability. Three items were adapted from Rudman et al., (2012) to measure likability. Participants rated: "How much do you like John/Jane?"; "Is John/Jane someone you want to get to know better?"; and "Would John/Jane be liked by people he/she associates with?" on seven-point Likert scales (1 = not at all; 7 = very much), α = .890.

Manipulation checks. This measure was designed to capture whether participants paid attention to the voluntary/involuntary manipulation. The first item consisted of the question: "Which of the following best describes John/Jane's procedure?" followed by four options: (1) John/Jane accidentally had his/her eye shape changed, (2) John/Jane chose to have his/her eye shape changed, (3) John/Jane accidentally had the shape of his/her nose changed, (4) John/Jane chose to have the shape of his/her nose changed. The second question was: "After the surgery, what eye shape did John/Jane have?" followed by two options: (1) Asian and, (2) White.

Results and Discussion

Cognitive and Affective Backlash

We predicted that a voluntary (versus accidental) change to the protagonist's racial phenotypic features would elicit cognitive and affective backlash in the form of vilification (i.e., a simultaneous exaggeration of a person's negative human uniqueness traits and underestimation of his/her positive human uniqueness traits) and dislike.

Vilification. We conducted a 2 (protagonist's race of origin: Asian versus white) by 3 (condition: control, voluntary change, and involuntary change) by 2 (valence: positive versus negative) mixed-factorial ANOVA on the uniquely human trait ratings. Overall, protagonists were perceived more positively than negatively, that is, there was a main effect of trait valence such that participants found all protagonists to possess higher levels of

positive traits ($M = 4.731$, $SE = .063$) than negative traits ($M = 2.893$, $SE = .071$), $F(1, 240) = 254.796$, $p < .001$, $\eta^2_p = .515$.

As predicted, there was a significant valence by condition interaction, $F(2, 240) = 17.273$, $p < .001$, $\eta^2_p = .126$ (Figure 1). Bonferroni adjusted simple effects analyses demonstrated that participants in the voluntary condition perceived the protagonist as possessing less positive uniquely human traits ($M = 4.268$, $SE = .109$) than both counterparts in the involuntary ($M = 4.953$, $SE = .106$), $t(240) = -3.672$, $p < .001$, $d = .568$, 95% CI [-1.051, -.318], and control conditions ($M = 4.972$, $SE = .111$), $t(240) = -3.689$, $p < .001$, $d = .583$, 95% CI [-1.078, -.329]. There was no significant difference between the involuntary condition and the control condition on positive trait ratings, $t(240) = .095$, $p = 1.000$, $d = .014$, 95% CI [-.388, .350].

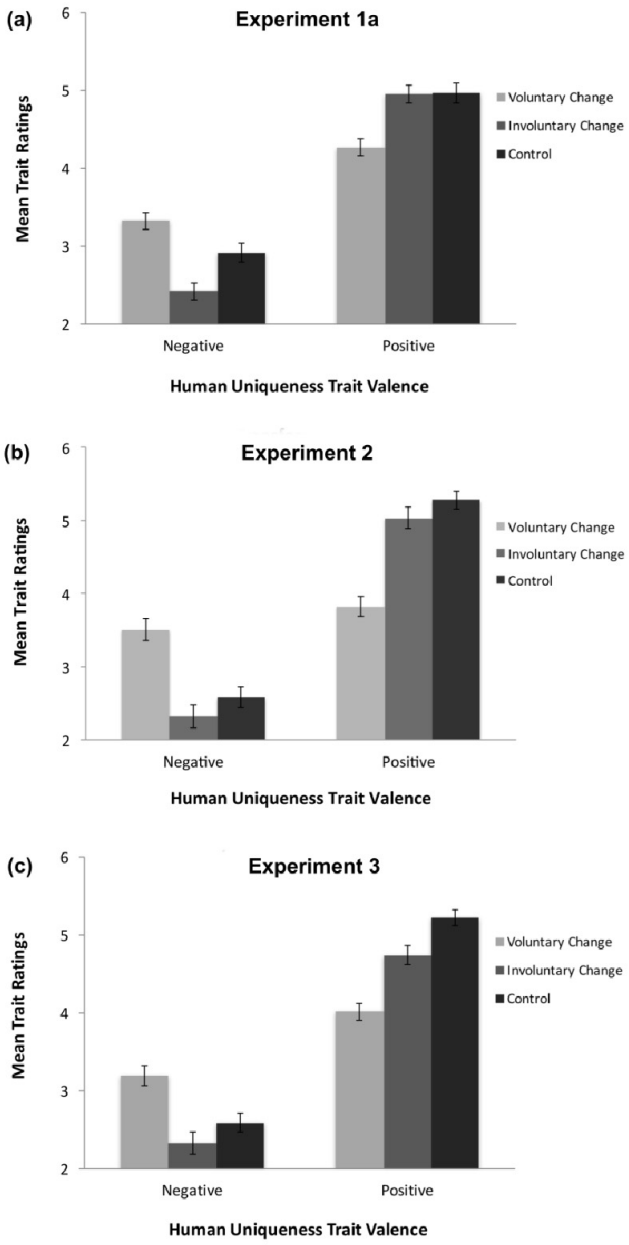

Figure 1. Vilification ratings by experimental condition in Experiment 1a (panel a), Experiment 2 (panel b), and Experiment 3 (panel c). Higher values indicate that the protagonist possesses the human uniqueness traits to a greater extent. Error bars represent +/- 1 standard error

Bonferroni adjusted simple effects analyses demonstrated that participants in the voluntary condition perceived the protagonist as having significantly and marginally significantly higher levels of negative human uniqueness traits ($M = 3.339$, $SE = .123$) compared to counterparts in the involuntary change ($M = 2.414$, $SE = .119$), $t(240) = 4.685$,

$p < .001$, $d = .725$, 95% CI [.513, 1.338], and control condition ($M = 2.925$, $SE = .124$), t (240) =2.053, $p = .056$, $d = .324$, 95% CI [-.007, .835], respectively.

Unexpectedly, participants in the involuntary condition deemed the protagonist to possess negative traits to a lesser extent than counterparts in the control condition, t (240) = -2.571, $p = .010$, $d = .400$, 95% CI [-.927, -.096]. This result might have indicated a "pity" effect for people who incur involuntary racial phenotypic change.

Taken together, these findings indicate that participants in the voluntary condition vilified the protagonist to a larger extent than counterparts in the involuntary and control conditions.

Likability. A 2 (protagonist's race of origin: Asian versus white) by 3 (condition: control, voluntary change, and involuntary change) between-subjects ANOVA was conducted on the 3-item composite measure of likability. As predicted, there was a significant effect of condition, F (2, 240) = 24.091, $p < .001$, $\eta^2_p = .167$ (Figure 2).

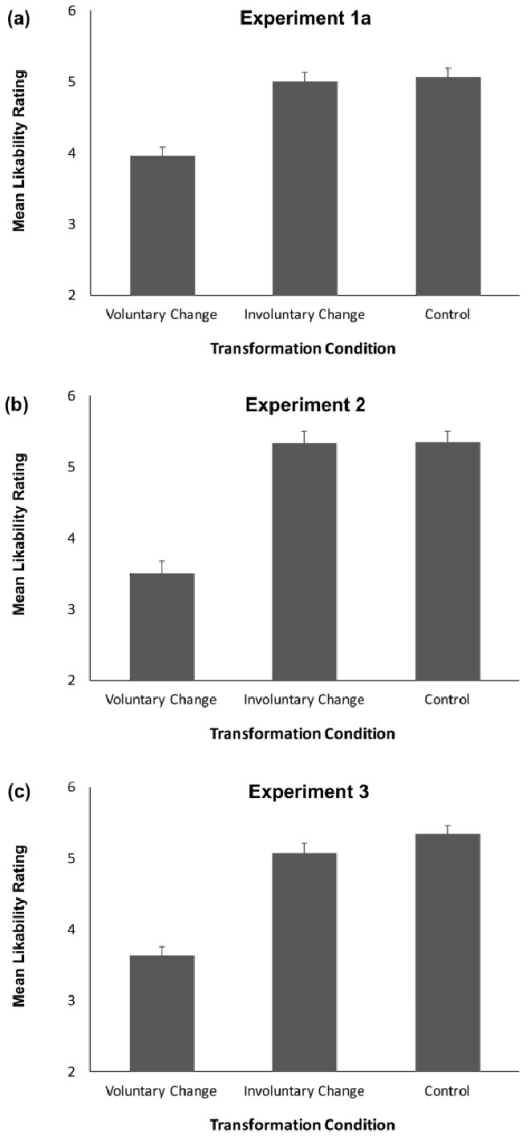

Figure 2. Likability by experimental condition in Experiment 1a (panel a), Experiment 2 (panel b), and Experiment 3 (panel c). Higher values are associated with greater likability of the protagonist. Error bars represent +/- 1 standard error

Bonferroni adjusted post-hoc comparisons indicated that, as predicted, participants in the voluntary condition liked the protagonist less ($M = 3.977$, $SE = .126$) as compared to counterparts in the involuntary ($M = 4.999$, $SE = .122$), t (240) = -5.841, $p < .001$, d =.904, 95% CI [-1.454, -.610] and control conditions ($M = 5.084$, $SE = .127$), t (240) =-6.195, $p < .001$, $d = .979$, 95% CI [-1.540, -.678]. There was no significant difference between the involuntary and control conditions, t (240) = -.482, $p = 1.000$, $d = .075$, 95% CI [-.501, .347].

Dissociation between Similarity and Categorization

We expected to find evidence for "origin essentialism" (Rips, 1989), such that post-phenotypic change, the protagonist will be judged as more similar to the new racial category but as belonging to the original racial category. To this end, we conducted a 2 (protagonist's race of origin: Asian versus white) by 3 (condition: control, voluntary change, and involuntary change) by 2 (rating type: categorization versus similarity) mixed-factorial ANOVA.

As predicted, there was a significant condition by rating type interaction, F (2, 240) = 12.430, $p < .001$, $\eta^2_p = .094$ (Figure 3). In the voluntary change condition, Bonferroni adjusted simple effects analyses showed that participants rated the protagonist's appearance as being more similar to the new racial category ($M = 2.038$, $SE = .104$), but that they categorized the protagonist's race as being more aligned with the protagonist's race of origin ($M = 1.480$, $SE = .090$), t (240) = 6.037, $p < .001$, $d = .948$, 95% CI [.376, .740]. This pattern was also observed in the involuntary condition, such that similarity ratings ($M = 2.021$, $SE = .101$) were higher than categorization ratings ($M = 1.419$, $SE = .087$), t (240) = 6.711, $p < .001$, $d = 1.023$, 95% CI [.425, .779].

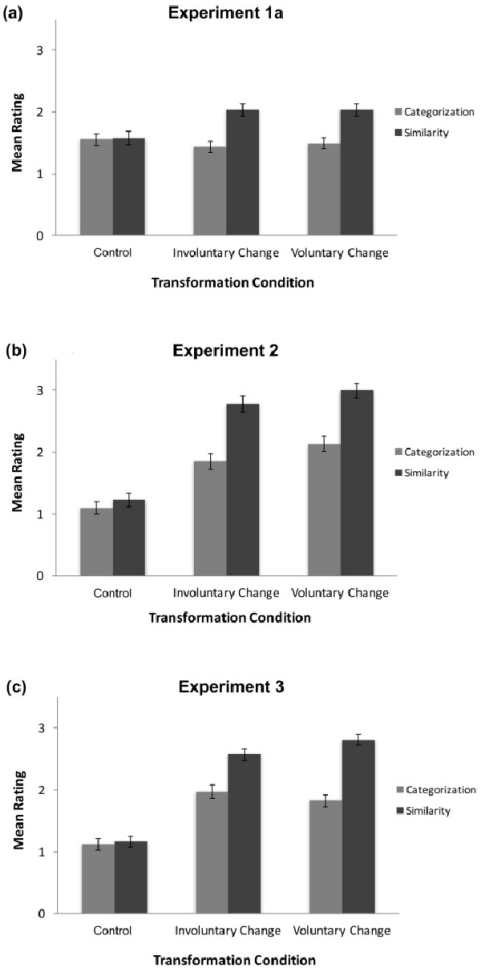

Figure 3. Dissociation between categorization and similarity ratings by experimental condition in Experiment 1a (panel a), Experiment 2 (panel b), and Experiment 3 (panel c). Lower values are associated with the protagonist's original category and higher values are associated with the protagonist's new category. Error bars represent +/- 1 standard error

In the control condition, the protagonist did not incur any change, so as expected there were no differences between similarity ($M = 1.565$, $SE = .105$) and categorization ($M = 1.550$, $SE = .091$) ratings, t (240) = .160, $p = .872$, $d = .025$, 95% CI [-.169, .199].

2.2 Experiment 1b

It is possible that Experiment 1a findings - that voluntary racial phenotypic change leads to backlash - resulted from the fact that the protagonist desired a phenotypic change, racial or otherwise. Thus, in Experiment 1b,

participants were given vignettes of white/Asian men/women who voluntarily changed a racially diagnostic feature (i.e., eye shape) or a race-neutral feature (i.e., chin shape). We predicted that a voluntary change to a racially diagnostic feature (versus a race-neutral) change would invoke greater cognitive and affective backlash in the form of vilification and dislike. Our investigation was designed to explore the global nature of social policing to voluntary phenotypic transgressors given that racial essentialism is foundational to human categorization (Prentice & Miller, 2007). We deemed it important, however, to select a similar number of Asian- and white-identified participants to prevent the findings from being biased by an unequal representation of participant race.

Method

Participants

One hundred participants were recruited via TurkPrime, a crowd-sourcing website that uses Amazon.com's MTurk. All participants had an MTurk approval rating of 90% or higher and lived in the United States. Participants received $.50 for their participation in the study. Seventeen participants were excluded (11 race-neutral condition, 6 racially diagnostic condition) for either incorrectly answering manipulation checks (described below), missing data, not identifying as Asian or white, or for not allowing their data to be used, leaving 83 participants (44 female, 39 male; 45 Asian-identified, 33 white-identified; M_{Age} = 33.2, SD_{Age} = 10.8) in the final analyses. Removing the 17 participants did not influence the results, so they were excluded from the final analysis.

Design

We employed a 2 (protagonist's race of origin: Asian versus white) x 2 (condition: racially diagnostic change versus race-neutral change) between-subjects factorial design.

Materials

Vignettes. The racially diagnostic change vignette was identical to the one used in Experiment 1a. The race-neutral change vignette was as follows:

John/Jane is a health conscious, Asian American/White man/woman in his/her early 30s with a successful career, a bright outlook on life, and no family history of serious illness. One day, John/Jane was in an accident that left a large, infected gash on his/her chin. He/She underwent surgery by an oral and maxillofacial doctor who was also a plastic surgeon. The surgeon gave John/Jane the option of keeping his/her chin the same or changing it. John/Jane requested that his/her chin be changed to appear narrower. The surgeon complied with his/her request and his/her chin ended up looking narrower.

Categorization vs. similarity dissociation measure. Identical to that used in Experiment 1a.

Vilification. Identical to that used in Experiment 1a. Positive items, α = .810, and negative items, α = .768.

Likability. Identical to that used in Experiment 1a, α = .919.

Manipulation checks. In all conditions, participants were asked, "Which of the following best describes John/Jane's procedure?" followed by four options: (1) John/Jane accidentally had his/her eye shape changed, (2) John/Jane chose to have his/her eye shape changed, (3) John/Jane accidentally had the shape of his/her chin changed, (4) John/Jane chose to have the shape of his/her changed. In the race-neutral change condition participants were asked: "After the surgery, what was the shape of John/Jane's chin?" followed by two options: (1) Wider, and (2) Narrower. Participants in the racially diagnostic change condition were given a second manipulation check identical to Experiment 1a's.

Procedure

After following the consent process approved by San Francisco State University's Institutional Review Board and agreeing to participate, participants were randomly assigned to one of the following experimental conditions: racially diagnostic change or race-neutral change. After reading the given vignette, participants completed the categorization versus similarity dissociation measure, provided vilification and likability ratings, and, finally, answered the manipulation checks and a basic demographics questionnaire.

Results and Discussion

Cognitive and Affective Backlash.

We predicted that a voluntary change to the protagonist's racial phenotypic features (versus a race-neutral feature) would elicit greater cognitive and affective backlash in the form of vilification (i.e., a simultaneous exaggeration of a person's negative human uniqueness traits and underestimation of his/her positive human uniqueness traits) and dislike.

Vilification. We conducted a 2 (protagonist's race of origin: Asian versus white) by 2 (condition: racially diagnostic change versus race-neutral change) by 2 (valence: positive versus negative) mixed-factorial ANOVA on the uniquely human trait ratings. Overall, protagonists were perceived more positively than negatively, that is, there was a main effect of trait valence such that participants found all protagonists to possess higher levels of positive traits (M = 4.375, SE = .106) than negative traits (M = 3.086, SE = .107), $F(1, 79) = 51.465$, $p < .001$, $\eta^2_p = .394$.

As predicted, there was a significant valence by condition interaction, $F(1, 79) = 31.888$, $p < .001$, $\eta^2_p = .288$. Bonferroni adjusted simple effects analyses demonstrated that participants in the racially diagnostic condition perceived the protagonist as possessing less positive uniquely human traits (M = 3.874, SE = .142) than counterparts in the race-neutral condition (M = 4.876, SE = .157), $t(79) = -3.972$, $p < .001$, d = .875, 95% CI [-1.423, -.582].

Additionally, participants in the racially diagnostic condition perceived the protagonist as possessing higher levels of negative traits (M = 3.600, SE = .143) compared to counterparts in the race-neutral condition (M = 2.573, SE = .158), $t(79) = 4.071$, $p < .001$, d = .896, 95% CI [.602, 1.451].

Likability. A 2 (protagonist's race of origin: Asian versus white) by 2 (condition: racially diagnostic change versus race-neutral change) between-subjects ANOVA was conducted on the 3-item composite measure of likability.

As predicted, there was a significant effect of condition, such that participants in the racially diagnostic condition liked the protagonist less (M = 3.506, SE = .185) than counterparts in the race-neutral condition (M = 4.692, SE = .204), $F(1,79) = 18.518$, $p < .001$, $\eta^2_p = .190$.

Dissociation between Categorization and Similarity

We expected to find evidence for "origin essentialism" (Rips, 1989), such that after altering a racially diagnostic feature, the protagonist will be judged as more similar to the new racial category but as belonging to the original racial category. To this end, we conducted a 2 (protagonist's race of origin: Asian versus white) by 2 (condition: racially diagnostic change versus race-neutral change) by 2 (rating type: categorization versus similarity) mixed-factorial ANOVA.

As predicted, there was a significant condition by rating type interaction, $F(1, 79) = 10.753$, $p = .002$, $\eta^2_p = .120$. In the racially diagnostic change condition, Bonferroni adjusted simple effects analyses showed that participants rated the protagonist's appearance as being more similar to the new racial category (M = 1.931, SE = .124), but that they categorized the protagonist's race as being more aligned with the protagonist's race of origin (M = 1.443, SE = .113), $t(79) = 5.565$, $p < .001$, d = 1.173, 95% CI [.313, .663]. The protagonist in the race-neutral condition did not incur any to racially diagnostic features, so as expected there were no differences between similarity (M = 1.378, SE = .137) and categorization ratings (M = 1.318, SE = .125), $t(79) = .628$, $p = .539$, d = .144, 95% CI [-.133, .252].

Experiment 1b provides some evidence that while there may be backlash towards targets who change racially diagnostic or neutral features, that perceivers tend to vilify targets who opt to change a racially diagnostic feature to a significantly larger extent than targets who choose to change a race-neutral feature.

2.3 Experiment 2

It is possible that findings from the previous experiments were stimuli driven, specifically by the given protagonist's race (i.e., Asian or white). Experiment 2 was designed to examine therefore whether Experiment 1a's findings would replicate in a different ethnic/racial context. To this end, we adapted the vignettes to depict Black or white protagonists who underwent surgery to alter their Afrocentric features (i.e., nose, lips, and skin tone, see Brown, Dane, & Durham, 1998). Similar to Experiment 1a, we predicted that a voluntary (versus accidental) change to the protagonist's features would invoke cognitive and affective backlash in the form of vilification and dislike (social policing). Our investigation was designed to explore the global nature of social policing to voluntary phenotypic transgressors given that racial essentialism is foundational to human categorization (Prentice & Miller, 2007). We deemed it important, however, to select a similar number of Black- and white-identified participants to prevent the findings from being biased by an unequal representation of participant race.

Method

Participants

Two hundred and thirty-four participants were recruited via TurkPrime, a crowd-sourcing website that uses Amazon.com's MTurk. All participants had an MTurk approval rating of 90% or higher and lived in the United

States. Participants received $.50 for their participation in the study. Thirty-four participants were excluded (20 involuntary condition, 14 voluntary condition, 0 control condition) for either incorrectly answering manipulation checks (described below), missing data, or for not allowing their data to be used, leaving 200 participants (141 female, 59 male; 105 Black-identified, 95 white-identified; M_{Age} = 33.6, SD_{Age} = 10.9) in the final analyses. Removing the 34 participants did not influence the results, so they were excluded from the final analysis.

Design

We employed a 2 (protagonist's race of origin: Black versus white) x 3 (Condition: control, voluntary, and involuntary) between-subjects factorial design.

Materials

Vignettes. Vignettes were adapted from Experiment 1a to depict Black and white protagonists.

Control.

John/Jane is a health-conscious, African American/White man/woman in his/her early 30s with a successful career, a bright outlook on life, and no family history of serious illness.

Involuntary versus voluntary change. The first part was identical to the control vignette. The second part was as follows:

John/Jane contracted meningitis and ended up 'losing his/her face,' such that his/her facial features became disfigured with severe damage to the skin. He/She underwent surgery by an oral and maxillofacial doctor who was also a plastic surgeon. The surgeon gave John/Jane the option of keeping his/her appearance the same or changing it.

Involuntary change. John/Jane requested that his/her facial features and skin tone stay the same. The surgeon tried to comply with his/her request, but due to a surgical error his/her facial features and skin tone ended up looking like a White/Black person's

Voluntary change. John/Jane requested that his/her facial features and skin tone be changed to look like a White/Black person's. The surgeon complied with his/her request and his/her facial features and skin tone ended up looking like a White/Black person's.

Categorization vs. similarity dissociation measure. Identical to that used in Experiments 1a and 1b. However, scale options were adapted for the current vignettes (1 = Black; 2 = More Black than White; 3 = More White than Black; 4 = White).

Vilification. Identical to that used in Experiments 1a and 1b. Positive items, α = .870, and negative items, α = .875.

Likability. Identical to that used in Experiments 1a and 1b, α = .894.

Manipulation checks. The first item consisted of the question: "Which of the following best describes John/Jane's procedure?" followed by four options: (1) John/Jane accidentally had his/her facial features and skin tone changed, (2) John/Jane chose to have his/her facial features and skin tone changed, (3) John's/Jane's procedure only changed his/her facial features, (4) John's/Jane's procedure only changed his/her skin tone. The second question was: "After the surgery, what did John/Jane's facial features and skin tone look like?" followed by two options: (1) A Black person's, (2) A White Person's.

Procedure

After following the consent process approved by San Francisco State University's Institutional Review Board and agreeing to participate, participants were randomly assigned to one of the experimental conditions: control, voluntary change, and involuntary change. After reading the given vignette, participants completed the categorization versus similarity dissociation measure, provided vilification and likability ratings, and, finally, answered the manipulation checks and a basic demographics questionnaire.

Results and Discussion

Cognitive and Affective Backlash

We predicted that a voluntary (versus accidental) change to the protagonist's racial phenotypic features would elicit cognitive and affective backlash in the form of vilification (i.e., a simultaneous exaggeration of a person's negative human uniqueness traits and underestimation of his/her positive human uniqueness traits) and dislike.

Vilification. We conducted a 2 (protagonist's race of origin: Black versus white) by 3 (condition: control, voluntary change, and involuntary change) by 2 (valence: positive versus negative) mixed-factorial ANOVA on the uniquely human trait ratings. Overall, protagonists were perceived more positively than negatively, that is,

there was a main effect of trait valence such that participants found all protagonists to possess higher levels of positive (M = 4.697, SE = .076) than negative traits (M = 2.807, SE = .088), F (1, 194) = 168.095, p < .001, η^2_p = .464.

As predicted, there was a significant valence by condition interaction, F (2, 194) = 23.693, p < .001, η^2_p = .196 (Figure 1). Bonferroni adjusted simple effects analyses demonstrated that participants in the voluntary condition perceived protagonists as possessing less positive uniquely human traits (M = 3.877, SE = .135) as compared to counterparts in the involuntary (M = 4.990, SE = .141), t (194) = -4.241, p < .001, d = .771, 95% CI = [-1.583, -.642], and control conditions (M = 5.224, SE = .121), t (194) = -5.530, p < .001, d = .934, 95% CI [-1.785, -.909]. There was no significant difference between the involuntary condition and the control condition on the positive trait ratings, t (194) = -.938, p = .624, d = .162, 95% CI [-.682, .214].

A similar pattern of backlash could be seen with the negative uniquely human traits. Participants in the voluntary condition perceived the protagonist as possessing higher levels of negative traits (M = 3.410, SE = .156) compared to counterparts in the involuntary (M = 2.423, SE = .162), t (194) = 3.761, p < .001, d = .684, 95% CI [.445, 1.530] and control conditions (M = 2.587, SE = .140), t (194) = 3.379, p < .001, d = .570, 95% CI [.318, 1.328]. There was no difference between the involuntary and control conditions on the negative trait ratings, t (194) = -0.657, p = 1.000, d = .113, 95% CI [-.681, .353].

Likability. A 2 (protagonist's race of origin: Black versus white) by 3 (condition: control, voluntary change, and involuntary change) between-subjects ANOVA was conducted on the 3-item composite measure of likability. As predicted, there was a significant effect of condition, F (2, 194) = 40.508, p < .001, η^2_p = .295 (Figure 2).

Bonferroni adjusted post-hoc comparisons indicated that perceivers in the voluntary condition liked the protagonist less (M = 3.554, SE = .162) than counterparts in the involuntary (M = 5.320, SE = .169), t (194) = -7.482, p < .001, d = 1.361, 95% CI [-2.317, -1.189] and control conditions (M = 5.328, SE = .145), t(194) = -8.181, p < .001, d = 1.381, 95% CI [-2.342, -1.295]. There was no significant difference between the involuntary and control conditions, t (194) = -.117, p = 1.000, d = .020, 95% CI [-.601, .470].

In sum, we found evidence (this time employing vignettes depicting Black and white protagonists) that a voluntary (versus accidental) change to the protagonist's racial phenotypic features invokes cognitive and affective backlash in the form of vilification and dislike (external policing).

Dissociation between Categorization and Similarity

We expected to find evidence for "origin essentialism" (Rips, 1989), such that post-phenotypic change, the protagonist will be judged as more similar to the new racial category but as belonging to the original racial category. To this end, we conducted a 2 (protagonist's race of origin: Black versus white) by 3 (condition: control, voluntary change, and involuntary change) by 2 (rating type: categorization versus similarity) mixed-factorial ANOVA.

As predicted, there was a significant condition by rating type interaction, F (2, 194) = 12.380, p < .001, η^2_p = .113 (Figure 3). In the voluntary change condition, Bonferroni adjusted simple effects analyses showed that participants rated the protagonist's appearance as being more similar to the new racial category (M = 2.879, SE = .113), but that they categorized the protagonist's race as being more aligned with the protagonist's race of origin (M = 2.088, SE = .115), t (194) = 6.489, p < .001, d = 1.156, 95% CI [.550, 1.032]. This pattern was also observed in the involuntary condition, such that similarity ratings (M = 2.739, SE = .118) were higher than categorization ratings (M = 1.842, SE = .120), t (194) = 7.061, p < .001, d = 1.311, 95% CI [.647, 1.148].

In the control condition, the protagonist did not incur any change, so as expected there were no differences between similarity (M = 1.268, SE = .102) and categorization (M = 1.120, SE = .103) ratings, t(194) = 1.359, p = .178, d = .216, 95% CI [-.068, .364].

2.4 Experiment 3

In alignment with the meaning maintenance model (see Heine, Proulx & Vohs, 2006), we have shown evidence that perceivers experience a meaning threat in response to learning that a target has chosen to undergo racial phenotypic change. In Experiment 3, we test this hypothesis directly by operationalizing a meaning threat as feelings of uncertainty about the fixity and predictability of 'race.' Our investigation was designed to explore the global nature of social policing to voluntary phenotypic transgressors given that racial essentialism is foundational to human categorization (Prentice & Miller, 2007). We deemed it important, however, to select a similar number of Black- and white-identified participants to prevent the findings from being biased by an unequal representation of participant race.

Method

Participants

Two hundred and ninety-five participants were recruited via TurkPrime, a crowd-sourcing website that uses Amazon.com's MTurk. All participants had an MTurk approval rating of 90% or higher and lived in the United States. Participants received $.50 for their participation in the study. Thirty-eight participants were excluded (23 involuntary condition, 11 voluntary condition, 4 control condition) for either incorrectly answering manipulation checks (described below), missing data, not identifying as either Black or white or for not allowing their data to be used, leaving 257 participants (179 female, 78 male; 131 Black-identified, 126 white-identified; M_{Age} = 35.1, SD_{Age} = 11.9) in the final analyses. Removing the 38 participants did not influence the results, so they were excluded from the final analysis.

Design

We employed a 2 (protagonist's race of origin: Black versus white) x 3 (Condition: control, voluntary, and involuntary) between-subjects factorial design.

Materials

Vignettes. Identical to those used in Experiment 2.

Categorization vs. similarity dissociation measure. Identical to that used in previous experiments.

Vilification. Identical to that used in previous experiments. Positive items, α = .862, and negative items, α = .889, traits.

Likability. Identical to that used in previous experiments, α = .882.

Uncertainty. This eight-item measure of uncertainty, which served as a proxy for a meaning threat, was designed to measure both the specific uncertainty elicited by John's/Jane's situation (e.g., "I dislike that John/Jane's situation could mean many different things") as well as a more general uncertainty about not knowing a target's race (e.g., "Not being able to identify one's race/ethnicity makes me feel anxious") on seven-point Likert scales (1 = absolutely untrue; 7 = absolutely true), α = .901.

Manipulation Checks. Identical to those used in Experiment 2.

Procedure

After following the consent process approved by San Francisco State University's Institutional Review Board and agreeing to participate, participants were randomly assigned to one of the experimental conditions: control, voluntary change, and involuntary change. After reading the given vignette, participants completed the categorization versus similarity dissociation measure and provided vilification and likability ratings. Participants in the involuntary and voluntary change conditions also completed the uncertainty measure and answered the manipulation checks. Finally, all participants completed a basic demographics questionnaire.

Results and Discussion

Cognitive and Affective Backlash

We predicted that a voluntary (versus accidental) change to the protagonist's racial phenotypic features would elicit cognitive and affective backlash in the form of vilification (i.e., a simultaneous exaggeration of a person's negative human uniqueness traits and underestimation of his/her positive human uniqueness traits) and dislike.

Vilification. We conducted a 2 (protagonist's race of origin: Black versus white) by 3 (condition: control, voluntary change, and involuntary change) by 2 (valence: positive versus negative) mixed-factorial ANOVA on the uniquely human trait ratings. Overall, protagonists were perceived more positively than negatively, that is, there was a main effect of trait valence such that participants found all protagonists to possess higher levels of positive (M = 4.661, SE = .066) than negative traits (M = 2.721, SE = .073), F (1, 251) = 259.944, $p < .001$, η^2_p = .509.

As predicted, there was a significant valence by condition interaction, F (2, 251) = 23.694, $p < .001$, η^2_p = .159 (Figure 1). Bonferroni adjusted simple effects analyses demonstrated that participants in the voluntary condition perceived the protagonist as possessing lower levels of positive traits (M = 4.025, SE = .111) than counterparts in the involuntary (M = 4.731, SE = .122), t (251) = -3.310, $p < .001$, d = .520, 95% CI [-1.103, -.310], and control conditions (M = 5.227, SE = .108), t (251) = -5.994, $p < .001$, d = .886, 95% CI [-1.575, -.831]. The protagonists who involuntarily changed their features were also rated as significantly lower on the positive traits than the control condition, t (251) = -2.354, p = .007, d = .365, 95% CI [-.888, -.105].

Negative uniquely human traits also replicated the findings of previous experiments. Participants in the voluntary condition perceived the protagonist as having significantly higher levels of negative traits (M = 3.221, SE = .124) than counterparts in the involuntary (M = 2.345, SE = .136), $t(251) = 4.107$, $p < .001$, d = .646, 95% CI [.432, 1.320] and control conditions (M = 2.597, SE = .121), t (251) = 3.124, p = .001, d = .460, 95% CI [.206, 1.041]. There was no difference between the involuntary and control conditions on the negative uniquely human trait ratings, t (251) = -1.196, p = .499, d = .185, 95% CI = [-.691, .186].

Likability. A 2 (protagonist's race of origin: Black versus white) by 3 (condition: control, voluntary change, and involuntary change) between-subjects ANOVA was conducted on the 3-item composite measure of likability. As in Experiment 2, we observed a significant effect of condition, F (2, 251) = 57.583, $p < .001$, $\eta^2_p = .315$ (Figure 2).

Bonferroni adjusted post-hoc comparisons indicated that, as predicted, participants in the voluntary condition liked the protagonist less (M = 3.617, SE = .122) as compared to counterparts in the involuntary (M = 5.056, SE = .134), $t(251) = -7.943$, $p < .001$, d = 1.249, 95% CI [-1.896, -1.023], and control conditions (M = 5.348, SE = .119), $t(251) = -10.163$, $p < .001$, d = 1.503, 95% CI [-2.136, -1.315]. There was no significant difference between the involuntary and control conditions, t (251) = -1.631, p = .415, d = .253, 95% CI [-.697, .165].

Meaning Threat. A 2 (protagonist's race of origin: Black versus white) by 2 (condition: voluntary change versus involuntary change) between-subjects ANOVA was conducted on the 8-item composite measure of uncertainty, which served as a proxy for a meaning threat. We observed a significant effect of condition, F (1, 159) = 60.203, $p < .001$, $\eta^2_p = .275$, such that participants in the voluntary condition experienced higher levels of uncertainty (M = 3.720, SE = .138), compared to participants in the involuntary condition (M = 2.134, SE = .151).

Dissociation Between Categorization and Similarity

We expected to find evidence for "origin essentialism" (Rips, 1989), such that post-phenotypic change, the protagonist will be judged as more similar to the new racial category but as belonging to the original racial category. To this end, we conducted a 2 (protagonist's race of origin: Black versus white) by 3 (condition: control, voluntary change, and involuntary change) by 2 (rating type: categorization versus similarity) mixed-factorial ANOVA.

As predicted, there was a significant condition by rating type interaction, F (2, 251) = 21.874, $p < .001$, $\eta^2_p = .148$ (Figure 3). In the voluntary change condition, Bonferroni adjusted simple effects analyses showed that participants rated the protagonist's appearance as being more similar to the new racial category (M = 2.822, SE = .089), but that they categorized the protagonist's race as being more aligned with the protagonist's race of origin (M = 1.826, SE = .098), t (251) = 9.600, $p < .001$, d = 1.431, 95% CI [.792, 1.201]. This pattern was also observed in the involuntary condition, such that similarity ratings (M = 2.595, SE = .098) were higher than categorization ratings (M = 2.003, SE = .108), t (251) = 5.203, $p < .001$, d = .855, 95% CI [.368, .816].

In the control condition, the protagonist did not incur any change, so as expected there were no differences between similarity (M = 1.161, SE = .087) and categorization (M = 1.118, SE = .096) ratings, t (251) = .425, p = .670, d = .062, 95% CI [-.156, .242].

3. General Discussion

The current set of studies provides evidence that perceivers react to a person who opts to change skin-deep but socially-laden racial phenotypic features by engaging in social policing of racial group boundaries: directing cognitive and affective backlash at the target (i.e., vilification and dislike, respectively). Maintaining 'race' as essentialized, versus conceiving of it as a socially constructed category (see Bodmer & Cavalli-Sforza, 1976; Molnar, 1992; Gould, 1981; Tate & Audette, 2001), provides cognitive economy (see Medin, 1989), a sense of meaning, order and predictability (see Haslam et al., 2000), as well as reinforces stereotypes (see Bastian & Haslam, 2006), which helps to keep this framework intact; a catch-22.

It is not the case, however, that people are born with an intuitive sense of 'race' (Hirschfeld, 1988; 1995). Instead, race becomes imbued with an alleged natural essence, in retrospect, when children learn that race matters in society (see Hirschfeld, 1995; Prentice & Miller, 2007; Rothbart & Taylor, 1992). Specifically, children begin sorting people into race categories by three to four years of age based on surface structural cues, such as skin tone (Katz, 1982; Davey, Mullin, Norburn & Pushkin, 1983; Ramsey, 1987). Around mid-childhood, racial thinking becomes theory-driven (Katz, 1982), and at this point and into adulthood, racial categories transform from being surface structural, or feature-based, into explanation-based concepts that entail causal thinking (Yuill, 1992). Skin tone, for example, is no longer just a cue that differentiates between people who belong to different social groups, but becomes "diagnostic" of a person's aggression and criminality, a window into an alleged causal essence (for racial phenotypicality bias, see Maddox, 2004; Maddox & Gray, 2002; for a skin tone memory bias, such that a

Black man appears lighter in the mind's eye following a counter-stereotypic prime, such as "educated," see Ben-Zeev et al., 2014). In sum, children notice racial surface-structural phenotypic differences early on but the causal meaning that these features become imbued with is a result of learned social beliefs.

Perceptions of causality between an alleged hidden binary and discrete essence (i.e., one that is or is not presumed to exist) and observable characteristics are foundational to inferences about the extent to which an exemplar is deemed to be a well-functioning category member. Rehder and Burnett (2005) demonstrated that when participants were introduced to a novel category with a causal structure - that is, in which a single feature was shown to cause all other features - participants perceived exemplars that possessed the causal feature but not any of its associated surface structural features as poor category exemplars (i.e., as less well-functioning). Our findings can be situated in a causal framework as follows: protagonists who opted to change their racial phenotypic features were likely subject to backlash because they chose to sever the "diagnostic" association between their perceived racial essence (i.e., causal feature) and their surface-structural phenotypic features; rendering themselves as lesser category members. The fact that protagonists in the voluntary conditions were judged as belonging to their racial category of origin, and thus as possessing a racial essence, was evidenced by the dissociation between similarity and categorization, which replicated across all three experiments.

The current study offers a foray into understanding the effects of voluntary racial phenotypic change as a window into how social policing of racial group boundaries serves to maintain essentialist beliefs about race. As such, it leaves some questions unresolved, and which beckon future investigations. First, given that racial essentialism is foundational to human categorization (e.g., Medin & Ortony, 1989; Prentice & Miller, 2007), the current study was designed to explore the global nature of social policing to voluntary racial phenotypic transgressors. Thus, we cared more about the "how," that is, the nature of backlash (i.e., vilification and dislike) directed at voluntary transgressors than the "why," or the content of perceivers' inferences regarding voluntary transgressors' motivations and perceivers' rationales for backlash and increased racial essentialism. Thus, the nature of perceivers' judgments remain unclear. It is possible that perceivers viewed protagonists of color who opted to change their racial phenotypic features to appear whiter as assimilating to Eurocentric norms (see, Davis, 2003; Haiken, 1997) whereas white protagonists who desired to appear more of color as appropriating marginalized outgroup norms (see, Brubaker, 2016). In any case, we advocate for future explorations designed to shed a more nuanced light on perceivers' specific rationales for engaging in social policing of racial phenotypic transgressors, including an examination of any intersectionality effects (e.g., potentially differential reasoning underlying punishing ingroup versus outgroup racial transgressors).

Second, one might argue that perhaps protagonists in the voluntary conditions were perceived as seeking to alter their racial identities, and if so, that this perceived desire for identity change (versus for racial cosmetic change alone) was responsible for backlash and increased essentialism. Inferences about appearance- versus identity-driven motivations are not likely to be perceived as orthogonal, however. Consider Davis's (2003, p.74) assertion that cosmetic surgery (racial or otherwise) can be seen as "…an intervention in identity rather than 'just' a beauty practice." That is, changing one's racial phenotypic features results in some degree of social passing, *de facto* - and thus in a somewhat novel social identity (see Ginsberg, 1996) - regardless of whether one's motivation was more aesthetically than identity driven to begin with. Thus, we contend that even if a protagonist would explicitly maintain their racial identity as intact, backlash would likely still occur.

Theory-based speculation aside, consider reactions to a celebrity of color who appears to have lightened his or her complexion substantially while still explicitly identifying as a person of color, such as the former baseball player Sammy Sosa. Perceivers' reactions tend to range from accusing the celebrity of self-loathing to accusing the celebrity of a complete rejection of his or her black identity (Hall, 2018). Thus, from both theory-based and cultural perspectives, it is reasonable to predict that even those individuals who choose to change their racial phenotypic features while maintaining their racial identities of origin, explicitly, would be subject to social policing. Nevertheless, the issue of whether and how racial appearance and identity motivated phenotypic change are linked and predict backlash provides rich fodder for future empirical investigations.

Third, the current voluntary and involuntary condition vignettes were designed to possess as much of an analogous structure as possible, which resulted in passive depictions of voluntary change. That is, in the voluntary condition, a protagonist did not actively seek racial phenotypic change but was offered this possibility as a secondary cosmetic procedure. In the future, it would be useful to explore whether employing vignettes that illustrate a protagonist's more active and direct voluntary pursuit of racial phenotypic change might elicit the same or perhaps even a greater degree of dislike and vilification.

Finally, the question of whether racial ingroups vs. outgroups might hold different reasons for engaging in vilification of and experiencing dislike to a racial transgressor is an open-ended one and rife for future investigations. Vilification and dislike might occur for different reasons and as a function of different moderators [e.g., social status, age, the extent to which a phenotypic change is deemed acceptable (e.g., straightening one's hair)], including whether a protagonist is a racial ingroup or outgroup member. Consider, yet again, the case of Sammy Sosa who has been vilified in the media by both Black and white people, but for seemingly different reasons. For Black perceivers (racial ingroups), Sammy Sosa's skin bleaching, and ensuing white appearance, seem to invoke judgments of racial betrayal, such as that of a "race traitor" or a person who experiences self-hatred and internalized racism (see, Hall, 2018; Nittle 2018). For example, Wendy Williams, a popular African American talk show host, exclaimed, "Sammy Sosa, …Wow, he really hated himself." (Nittle, 2018). In an op-ed, Hall (2018), argued that when famous Black people whiten their racial phenotypic features, including the "king of pop" Michael Jackson, Sammy Sosa and rapper Nicki Minaj, these celebrities are met with accusations of not being "black enough."

White perceivers' (racial outgroups) vilification and dislike narratives seem to be characterized by negative judgments of a different flavor. Some white perceivers, for example, deem people of color's attempts to whiten their looks, including Sammy Sosa's, to be indicative of poor mental health (see, Williams, 2017). The similarities and differences between ingroup and outgroup members' vilification and dislike narratives lie outside the scope of the current work and constitute a worthwhile endeavor for future investigations. A fruitful direction might be to situate such investigations in the growing literature on "acting white" accusations and cultural invalidations - purposeful or inadvertent threats to an individual's social identity - by racial ingroups and outgroups (see, Durkee, Gazley, Hope & Keels, 2019).

Overall, the present findings offer some evidence that perceivers respond to a target's voluntary racial phenotypic change by policing racial group boundaries - vilifying and disliking the target (cognitive and affective backlash). This essentialized view of race offers a way to make sense of the world while buffering existential angst (see Soloman, Greenberg & Pyszczynski, 1991). However, racial essentialism is not only counter-scientific but has also been documented to cause detrimental outcomes, such as stereotyping and discrimination (Bastian & Haslam, 2006), and the maintenance of a hierarchical racial structure that privileges certain groups and discriminates against others (Smedley & Smedley, 2005). It thus behooves us as a society to heed Hirschfeld's (1998) exhortation, "It is precisely because race is essentialized that it serves systems of power and authority" (p.73), by helping to shift social views to embrace race as a social construction - advancing multiculturalism (e.g., Plaut, Thomas & Goran, 2009) while eschewing colorblindness (Neville, Lilly, Duran, Lee, & Browne, 2000) - in a way that celebrates racial group differences, and without racism.

References

Bastian, B., & Haslam, N. (2006). Psychological essentialism and stereotype endorsement. *Journal of Experimental Social Psychology, 42*, 228-235. https://doi.org/10.1016/j.jesp.2005.03.003

Bastian, B., & Haslam, N. (2010). Excluded from humanity: The dehumanizing effects of social ostracism. *Journal of Experimental Social Psychology, 46*, 107-113. https://doi.org/10.1016/j.jesp.2009.06.022

Ben-Zeev, A., Dennehy, T. C., Goodrich, R. I., Kolarik, B. S., & Geisler, M. W. (2014). When an "educated" Black man becomes lighter in the mind's eye: Evidence for a skin tone memory bias. *SAGE Open, 4*, 1-9. https://doi.org/10.1177/2158244013516770

Bodmer, W. F., & Cavalli-Sforza, L. L. (1976). *Genetics, evolution, and man* (pp. 231-258). San Francisco, CA: WH Freeman. https://doi.org/10.1002/ajpa.1330390117

Brown, T. D., Dane, F. C., & Durham, M. D. (1998). Perception of race and ethnicity. *Journal of Social Behavior and Personality, 13*, 295-306.

Brubaker, R. (2016). The Dolezal affair: Race, gender, and the micropolitics of identity. *Ethnic and Racial Studies, 39*, 414-448. https://doi.org/10.1080/01419870.2015.1084430

Davey, A., Mullin, P. N., Norburn, M. V., & Pushkin, I. (1983). *Learning to be prejudiced: Growing up in multi-ethnic Britain*. USA: E. Arnold.

Davis, K. (2003). Surgical passing: or why Michael Jackson's nose makes us' uneasy. *Feminist Theory, 4*, 73-92. https://doi.org/10.1177/1464700103004001004

Doland, A. (2014). Forger 'Real Beauty': Ads for skin-whitening beauty products just won't die. *Adage*. Retrieved from http://adage.com/article/global-news/awful-ads-skin-whitening-products-die/294766/

Dreisinger, B. (2008). *Near Black: White-to-Black passing in American culture*. Amherst, MA: University of Massachusetts Press.

Durkee, M. I., Gazley, E. R., Hope, E. C., & Keels, M. (2019). Cultural Invalidations: Deconstructing the "Acting White" Phenomenon Among Black and Latinx College Students. *Cultural Diversity and Ethnic Minority Psychology*. https://doi.org/10.1037/cdp0000288

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*, 175-191. https://doi.org/10.3758/BF03193146

Fincher, K. M., & Tetlock, P. E. (2016). Perceptual dehumanization of faces is activated by norm violations and facilitates norm enforcement. *Journal of Experimental Psychology: General, 145*, 131-146. https://doi.org/10.1037/xge0000132

Fiske, S. T. (2002). What we know now about bias and intergroup conflict, the problem of the century. *Current Directions in Psychological Science, 11*, 123-128. https://doi.org/10.1111/1467-8721.00183

Gelman, S. A. (2003). *The essential child: Origins of essentialism in everyday thought*. New York, NY: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195154061.001.0001

Gould, S. J. (1981). *The mismeasure of man*. New York, NY: Norton. https://doi.org/10.1002/1520-6807(198410)21:4<528::AID-PITS2310210423>3.0.CO;2-Y

Haiken, E. (1997). *Venus envy: A history of cosmetic surgery*. Baltimore, MD: Johns Hopkins University Press.

Hall, R. (2018). Black America's 'bleaching syndrome'. *The Conversation*. Retrieved from http://theconversation.com/black-americas-bleaching-syndrome-82200

Haslam, N. (2006). Dehumanization: An integrative review. *Personality and Social Psychology Review, 10*, 252-264. https://doi.org/10.1207/s15327957pspr1003_4

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

Heine, S. J., Proulx, T., & Vohs, K. D. (2006). The meaning maintenance model: On the coherence of social motivations. *Personality and Social Psychology Review, 10*, 88-110. https://doi.org/10.1207/s15327957pspr1002_1

Heyes, C. J. (2006). Changing race, changing sex: The ethics of self-transformation. *Journal of Social Philosophy, 37*, 266-282. https://doi.org/10.1111/j.1467-9833.2006.00332.x

Hickman, C. B. (1997). The devil and the one drop rule: Racial categories, African Americans, and the US census. *Michigan Law Review, 95*, 1161-1265. https://doi.org/10.2307/1290008

Hirschfeld, L. A. (1988). On acquiring social categories: Cognitive development and anthropological wisdom. *Man*, 611-638. https://doi.org/10.2307/2802596

Hirschfeld, L. A. (1995). The inheritability of identity: Children's understanding of the cultural biology of race. *Child Development, 66*, 1418-1437. https://doi.org/10.2307/1131655

Hirschfeld, L. A. (1998). *Race in the making: Cognition, culture, and the child's construction of human kinds*. Cambridge, MA: MIT Press. https://doi.org/10.7551/mitpress/5734.001.0001

Hunter, M. L. (2005). *Race Gender and the Politics of Skin Tone*. New York, NY: Routledge. https://doi.org/10.4324/9780203620342

Hunter, M. L. (2007). The Persistent Problem of Colorism: Skin Tone, Status, and Inequality. *Sociology Compass, 1*, 237-254. https://doi.org/10.1111/j.1751-9020.2007.00006.x

Hunter, M. L. (2011). Buying racial capital: Skin-bleaching and cosmetic surgery in a globalized world. *The Journal of Pan African Studies, 4*, 142-164.

Jost, J. T., & Banaji, M. R. (1994). The role of stereotyping in system-justification and the production of false consciousness. *British Journal of Social Psychology, 33*, 1-27. https://doi.org/10.1111/j.2044-8309.1994.tb01008.x

Katz, P. A. (1981). Development of children's racial awareness and intergroup attitudes. In L.G. Katz (Eds.), *Current topics in early childhood education* (Vol. 4, pp. 16-54). New York, NY: Ablex.

Kruglanski, A. W., & Webster, D. M. (1996). Motivated closing of the mind: "Seizing" and "freezing." *Psychological Review, 103*, 263. https://doi.org/10.1037/0033-295X.103.2.263

Leyens, J. P., Rodriguez-Perez, A., Rodriguez-Torres, R., Gaunt, R., Paladino, M. P., Vaes, J., & Demoulin, S. (2001). Psychological essentialism and the differential attribution of uniquely human emotions to ingroups and outgroups. *European Journal of Social Psychology, 31*, 395-411. https://doi.org/10.1002/ejsp.50

Maddox, K. B. (2004). Perspectives on racial phenotypicality bias. *Personality and Social Psychology Review, 8*, 383-401. https://doi.org/10.1207/s15327957pspr0804_4

Maddox, K. B., & Gray, S. A. (2002). Cognitive representations of Black Americans: Reexploring the role of skin tone. *Personality and Social Psychology Bulletin, 28*, 250-259. https://doi.org/10.1177/0146167202282010

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

Medin, D. L., & Ortony, A. (1989). Psychological essentialism. *Similarity and analogical reasoning*, 179-195. https://doi.org/10.1017/CBO9780511529863.009

Millner, D. (2017, March 3). Why Rachel Dolezal can never be Black. *National Public Radio*. Retrieved from http://www.npr.org/sections/codeswitch/2017/03/03/518184030/why-rachel-dolezal-can-never-be-black?sc=tw

Molnar, S. (1992). *Human variation: Races, types, and ethnic groups*. Englewood Cliffs, NJ: Prentice Hall. https://doi.org/10.1002/ajpa.1330630413

Neville, H. A., Lilly, R. L., Duran, G., Lee, R. M., & Browne, L. (2000). Construction and initial validation of the color-blind racial attitudes scale (CoBRAS). *Journal of Counseling Psychology, 47*, 59-70. https://doi.org/10.1037/0022-0167.47.1.59

Nittle, N. (2018). Sammy Sosa Is a victim of colorism. *Racked*. Retrieved from https://www.racked.com/2018/2/15/17013740/sammy-sosa-bleached-skin-colorism-amara-la-negra

Ono, Y., (1966). White chess set [Acrylic on wood]. Museum of Modern Art, New York, NY.

Plaut, V. C., Thomas, K. M., & Goren, M. J. (2009). Is multiculturalism or color blindness better for minorities? *Psychological Science, 20*, 444-446. https://doi.org/10.1111/j.1467-9280.2009.02318.x

Prentice, D. A., & Miller, D. T. (2007). Psychological essentialism of human categories. *Current Directions in Psychological Science, 16*, 202-206. https://doi.org/10.1111/j.1467-8721.2007.00504.x

Proulx, T., & Heine, S. J. (2008). The case of the transmogrifying experimenter affirmation of a moral schema following implicit change detection. *Psychological Science, 19*, 1294-1300. https://doi.org/10.1111/j.1467-9280.2008.02238.x

Proulx, T., & Heine, S. J. (2010). The frog in Kierkegaard's beer: Finding meaning in the threat-compensation literature. *Social and Personality Psychology Compass, 4*, 889-905. https://doi.org/10.1111/j.1751-9004.2010.00304.x

Ramsey, P. G. (1987). Young children's thinking about ethnic differences. In J. Phinney & M. Rotheram (Eds.), *Children's ethnic socialization: Pluralism and development*, 56-72. London: Sage Publications. https://doi.org/10.1080/15240750802432607

Rehder, B., & Burnett, R. C. (2005). Feature inference and the causal structure of categories. *Cognitive Psychology, 50*, 264-314. https://doi.org/10.1016/j.cogpsych.2004.09.002

Rips, L. J. (1989). Similarity, typicality, and categorization. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 21-59). USA: Cambridge University Press. https://doi.org/10.1017/CBO9780511529863.004

Rothbart, M., & Taylor, M. (1992). Category labels and social reality: Do we view social categories as natural kinds?. In G. Semin & K. Fiedler (Eds.), *Language, interaction and social cognition* (pp. 11-36). London, England: Sage.

Rudman, L. A. (1998). Self-promotion as a risk factor for women: The costs and benefits of counterstereotypical impression management. *Journal of Personality and Social Psychology, 74*, 629-645. https://doi.org/10.1037/0022-3514.74.3.629

Rudman, L. A., & Fairchild, K. (2004). Reactions to counterstereotypic behavior: The role of backlash in cultural stereotype maintenance. *Journal of Personality and Social Psychology, 87*, 157-176. https://doi.org/10.1037/0022-3514.87.2.157

Rudman, L. A., & Glick, P. (2001). Prescriptive gender stereotypes and backlash toward agentic women. *Journal of Social Issues, 57*, 743-762. https://doi.org/10.1111/0022-4537.00239

Rudman, L. A., & Phelan, J. E. (2008). Backlash effects for disconfirming gender stereotypes in organizations. *Research in Organizational Behavior, 28*, 61-79. https://doi.org/10.1016/j.riob.2008.04.003

Rudman, L. A., Moss-Racusin, C. A., Phelan, J. E., & Nauts, S. (2012). Status incongruity and backlash effects: Defending the gender hierarchy motivates prejudice against female leaders. *Journal of Experimental Social Psychology, 48*, 165-179. https://doi.org/10.1016/j.jesp.2011.10.008

Smedley, A., & Smedley, B. D. (2005). Race as biology is fiction, racism as a social problem is real: Anthropological and historical perspectives on the social construction of race. *American Psychologist, 60*, 16-26. https://doi.org/10.1037/0003-066X.60.1.16

Solomon, S., Greenberg, J., & Pyszczynski, T. (1991). A terror management theory of social behavior: The psychological functions of self-esteem and cultural worldviews. *Advances in Experimental Social Psychology, 24*, 93-159. https://doi.org/10.1016/S0065-2601(08)60328-7

Srivastava, S., & Beer, J. S. (2005). How self-evaluations relate to being liked by others: Integrating sociometer and attachment perspectives. *Journal of Personality and Social Psychology, 89*, 966-977. https://doi.org/10.1037/0022-3514.89.6.966

Tate, C., & Audette, D. (2001). Theory and research on 'race' as a natural kind variable in psychology. *Theory & Psychology, 11*, 495-520. https://doi.org/10.1177/0959354301114005

Wakslak, C. J., Jost, J. T., & Bauer, P. (2011). Spreading rationalization: Increased support for large-scale and small-scale social systems following system threat. *Social Cognition, 29*, 288-302. https://doi.org/10.1521/soco.2011.29.3.288

Wald, G. (2000). *Crossing the line: Racial passing in twentieth-century US literature and culture*. Durham, NC: Duke University Press. https://doi.org/10.2307/j.ctv11smxsg

Williams, J. (2017). Sammy Sosa's white skin has Twitter freaking out (again). *Newsweek*. Retrieved from https://www.newsweek.com/sammy-sosa-skin-white-2017-704457

Yuill, N. (1992). Children's conception of personality traits. *Human Development, 35*, 265-279. https://doi.org/10.1159/000277220