Journal of Experimental Social Psychology 102 (2022) 104348

Contents lists available at ScienceDirect

Journal of Experimental Social Psychology

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

Equating silence with violence: When White Americans feel threatened by anti-racist messages☆

Frank J. Kachanoff[a,*], Nour Kteily[b], Kurt Gray[c]

[a] Wilfrid Laurier University
[b] Northwestern University
[c] The University of North Carolina at Chapel Hill

ARTICLE INFO

Keywords:
Structural racism
Anti-racism education
Social identity threat
Racial and ethnic attitudes and relations

ABSTRACT

Anti-racist messages educate people about structural racism and argue that indifference and inaction are the foundational building-blocks of race-based inequities. But these messages generate backlash, with several American states banning education about structural racism. We hypothesized that White Americans experience White identity threat and resist anti-racist messages most when they interpret these messages to *equalize* a lack of anti-racist action (i.e., indifference and silence), treating it as though it were the *same as* blatant racism. In contrast, we predicted that interpreting anti-racist messages to position silence as a foundational "building-block" for blatant racism would not evoke backlash. In Study 1 (N = 428) ~55% of White respondents in a representative American sample interpreted anti-racist messages as equating indifference with violence, and an equalizing interpretation predicted White identity threat and message resistance. In Study 2 (N = 492) we found that experimentally manipulating anti-racist messages to evoke high vs. low levels of equalizing interpretation led White Americans to feel more White identity threat and in turn be more resistant to both the anti-racist message and anti-racist action in general. In Study 3 (N = 1337) seeing anti-racist messages (vs. no-message) had little effect on White Americans in general, but evoked identity threat and denial of racism among White Americans high in equalizing interpretation who did not interpret the messages as conveying inaction to be a building-block for structural racism. In Study 4a and 4b (N = 789), we reveal a successful nudge for making anti-racist messages less threatening and more motivating for White Americans by using language less likely to evoke an equalizing interpretation.

Our understanding of racism is changing. Everyday people and psychologists have long tied racism to the actions and attitudes—-whether explicit (Allport, 1954; Dovidio & Gaertner, 1986) or implicit (Greenwald & Banaji, 1995; Sue, 2010)—of individual "bad apples" (Asare, 2020; Gillborn, 2006; Schmidt, 2005). However, the COVID-19 pandemic's disparate impact on people of color, coupled with the continued murders of unarmed Black people by police has challenged White people in the United Sates (and globally) to recognize that *structural racism*—the different behaviors, history, policies, and institutions that give rise to racial inequities (Ansley, 1997; Salter, Adams, & Perez, 2018) — is still imbedded in our society (Ford, Green, & Gross, 2022; Ledgerwood et al., 2022).

Critical race theory (CRT; Ansley, 1997) argues that, to attenuate racial inequities, it is not sufficient for White people to focus solely on regulating their own racial biases (not being racist); rather, they must also actively work to challenge race-based structural inequities (being ***anti***-racist; Kendi, 2019). Anti-racist messages (inspired by or compatible with CRT) often argue for the role of White silence in maintaining systems of racial inequity (either in a picture or a short phrase)—but these messages are divisive and have received national attention including on Fox News and CBS (Fox News, Jan 18th, 2018; Capatides, 2020), on social media platforms (Capatides, 2020), and within organizations and schools (Fox News, Jan 18th, 2018; Pothast, 2021). Several US states (e.g., Florida, Iowa, and New Hampshire; Ray & Gibbons, 2021) and private organizations (Pothast, 2021) have even banned discussion about race altogether. Here, we examine the psychological roots of this division: We suggest that White Americans oppose anti-racist messages most when they interpret them to equate silence to be

☆ This paper has been recommended for acceptance by Jarret Crawford.
* Corresponding author.
E-mail address: fkachanoff@wlu.ca (F.J. Kachanoff).

https://doi.org/10.1016/j.jesp.2022.104348
Received 9 October 2021; Received in revised form 30 April 2022; Accepted 4 May 2022
Available online 21 June 2022

the same as violence (i.e., an *equalizing interpretation*) rather than a foundational building-block for race-based violence and structural inequities to persist in society (i.e., a *building-block interpretation*). Ultimately, we test whether modifying anti-racist messages to be less likely to evoke an equalizing interpretation can minimize backlash without watering down their core message that White silence is a building-block for structural inequities and violence.

1. Anti-racist messages about structural racism

Critical race theory (CRT; Ansley, 1997) is a set of ideas about the legacy and present manifestation of racism in the United States (Ray & Gibbons, 2021). CRT argues that structural racial inequities and racial biases existing today stem from racial power differences that originated during the transatlantic slave trade and European colonialism (Ansley, 1997). In this way, CRT describes racism as a top-down process where power differences tied to racialized identities created hundreds of years ago continue to shape people's racialized experiences today. Importantly, CRT also argues that dismantling structural racism requires the bottom-up process of privileged group members working to dismantle racist systems: this can be achieved when White people reflect on how their attitudes, behavior, and privilege is influenced by racialized power inequities, and actively challenge ongoing inequities (see Kendi, 2019 for review).

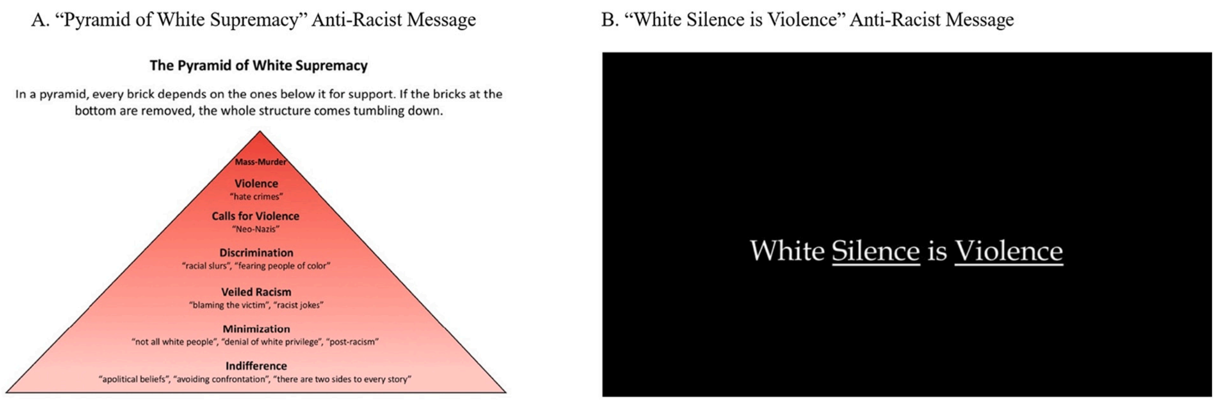
The top-down and bottom-up processes emphasized by CRT are captured in anti-racist messages. One prominent—and controversial—example is the Pyramid of White Supremacy (PWS; sosspeace.org, 2019; see Fig. 1) which argues that different layers of a racist system can be represented structurally as a pyramid. The top layers of the pyramid contain active violence and inequities imbedded into social institutions (e.g., police brutality and hate crimes). Racist jokes and more subtle insensitivities like microaggressions (Williams, 2020) are contained in the middle. Finally, denial and indifference to active racism and structural inequities are illustrated at the Pyramid's base. All levels of the pyramid reinforce each other and represent racism as a structural process. Thus, a core message of the pyramid is that only when White individuals actively speak up against all forms of racism (individual and structural)— destabilizing the foundation of the pyramid—will the entire racist system (i.e., the pyramid) crumble. A similar idea is intended to be conveyed by the message "White silence is violence" which emerged during the Black Lives Matter movement. While briefer than the PWS, the White silence message similarly argues that if White Americans are silent about racial issues, then racial violence and inequities will persist (Capatides, 2020). In short, anti-racist messages argue that racist structures are either maintained or toppled depending on whether people stay silent or act anti-racist (Kendi, 2019).

2. The benefits and barriers of White Americans confronting structural racism

How do White Americans react to real-world anti-racist messages? This question is currently unexplored—a void we fill in the present work—but there is past work that examines what happens when White people are encouraged to confront structural racism. Educating White Americans to think about racism as a structural (versus strictly individual) process has the potential to increase White American's support for attenuating racial inequities (Adams, Edkins, Lacka, Pickett, & Cheryan, 2008; Bonam, Nair Das, Coleman, & Salter, 2019; Rucker, Duker, & Richeson, 2019; Rucker & Richeson, 2021). University students who are taught about structural *and* individual racism (versus *only* about individual racism) show more support for social policies that redistribute power equally across different racial groups (e.g., affirmative action; Adams et al., 2008). Similarly, White Americans who conceive of racism as structural (versus interpersonal) more accurately detect and support reducing inequities that exist in the criminal justice system (Rucker et al., 2019; Rucker & Richeson, 2021). Yet while these studies underscore the importance of having White Americans think about racism as a structural process, other research suggests that many White Americans might resist this idea.

White Americans often find it psychologically challenging to confront the privileges they experience because of structural inequities (Ford et al., 2022; Knowles, Lowery, Chow, & Unzueta, 2014; Phillips & Lowery, 2015; Takahashi & Jefferson, 2021; Unzueta & Lowery, 2008). White people experience moral threat when confronted with structural racism (Knowles et al., 2014; Lowery, Knowles, & Unzueta, 2007; Unzueta & Lowery, 2008) and feel powerless when talking about racism (Takahashi & Jefferson, 2021). The identity-based threat and ensuing negative emotions that White Americans experience when confronting structural racism is referred to in some quarters as a 'White fragility' response (Ford et al., 2022; DiAngelo, 2018). Ultimately, while some White Americans might deny the existence of structural racism to alleviate their threat, others might seek to dismantle racist structures (Knowles et al., 2014).

No work has examined how White Americans react to real-world anti-racist messages about structural racism, especially when encountering these messages without a trained anti-racism educator to guide their interpretation (see Tatum, 1992 for a review of White people's responses to formal anti-racism education). Moreover, while past work suggests that White Americans may be generally threatened by the idea of structural racism, we know less about which *type* of White Americans become more threated than others. Here, we examine how White people respond to anti-racist messages about structural racism, of the form that regularly appear on the evening news, on social media, and informally within organizations—all part of an ongoing national dialogue about racism (Capatides, 2020; Pothast, 2021). We suggest that White people's



A. "Pyramid of White Supremacy" Anti-Racist Message

The Pyramid of White Supremacy

In a pyramid, every brick depends on the ones below it for support. If the bricks at the bottom are removed, the whole structure comes tumbling down.

Mass-Murder

Violence
"hate crimes"

Calls for Violence
"Neo-Nazis"

Discrimination
"racial slurs", "fearing people of color"

Veiled Racism
"blaming the victim", "racist jokes"

Minimization
"not all white people", "denial of white privilege", "post-racism"

Indifference
"apolitical beliefs", "avoiding confrontation", "there are two sides to every story"

B. "White Silence is Violence" Anti-Racist Message

White Silence is Violence

Fig. 1. Standard "pyramid of white supremacy" illustration and white silence is violence message shown in Study 1.

responses to these messages will be shaped by individual differences in how they interpret them.

3. An equalizing interpretation of anti-racist messages and its consequences

Although some White Americans support anti-racist messages about structural racism (and CRT more broadly) others are resistant (Ray & Gibbons, 2021). For example, in 2021, the tech company Basecamp saw about one-third of its workforce quit after becoming embroiled in controversy over issues relating to an anti-racist message. The conflict started when an employee shared the "pyramid of hate" (similar to the PWS) in Basecamp's internal forum, as part of an apology for having participated in a company practice of keeping a list of "funniest-sounding customer names." The employee realized that this practice could help contribute to structural racism, but other employees were offended at the notion that their behavior could in any way be contributing to white supremacy or extreme acts like genocide atop of the pyramid. In light of the conflict, Basecamp's upper management moved to ban all discourse of race and politics, contributing to the employee exodus (Pothast, 2021).

Why this division? We suggest that anti-racist messages about structural racism are divisive because people vary in how much they endorse two different interpretations of these messages—an equalizing interpretation and a building-block-interpretation. We hypothesize that those who support messages like the PWS see them as advancing a "building block" interpretation, where indifference about race issues creates an environment—both psychologically and socially—that ultimately allows for racial inequities and racial violence to continue uninterrupted.

On the other hand, we hypothesize that those who oppose these messages see them as advancing an "equalizing" message, whereby inaction about racism is judged as equally harmful/immoral as blatant racism (e.g., "not challenging racist jokes is the same as racial violence"). For instance, executive David Heinemeier Hansson of Basecamp opposed the pyramid of hate being shared in his company's internal forum because he was concerned that it equated insensitive jokes with colonial oppression. He stated: "we can recognize that forceful renaming by a colonial regime is racist and wrong while also recognizing that having a laugh at customer names behind their back is inappropriate and wrong without equating or linking the two" (Hansson, April 28th 2021, personal blogpost). But supporters of anti-racism messages, like Emily Pothast (a historian who contributed a Medium article describing the Basecamp controversy), did not see these same messages through an equalizing lens (Pothast, 2021). Pothast suggested that messages like the pyramid do not "equate making fun of names with genocide" but rather "demonstrate that hate crimes and structural racism don't happen in a vacuum. They happen within a society in which a foundation has been laid that makes them possible" (Pothast, 2021).

One reason why equalizing interpretations might generate backlash among White Americans is threat to their group's moral identity (Gunn & Wilson, 2011). Past research shows that in general White Americans are more threatened by structural vs. individual definitions of racism (Unzueta & Lowery, 2008). It is easier for White Americans to live with the idea of ongoing racism when they can cast the blame on the behavior of a "few bad apples," rather than the social systems they might help sustain and benefit from (Wetherell & Potter, 1992). But the idea of structural racism may be *especially* threatening to White Americans when they have an equalizing interpretation. When people hold equalizing interpretations, they not only have to reconcile that members of their group might help maintain racist systems, they assume that committing more mild acts—like failing to confront a work colleague's racist joke—makes them as morally "evil" as someone who commits genocide.

Beyond the simple fact that no one likes being labeled a racist, holding equalizing interpretations might threaten White Americans' perceived collective autonomy—their perceived freedom to express their group identity (Kachanoff, Kteily, Khullar, Park, & Taylor, 2020). In the current societal environment where there are initiatives to remove monuments of White historical figures with ties to racism (McGivney, 2021), some White Americans might fear that expressing Whiteness will be perceived as supporting white supremacy. An equalizing interpretation likely exacerbates collective autonomy threat by implying that White identity expression is equally as immoral as blatant racism.

Although guilt about past injustices can motivate groups to address those injustices (Shnabel & Nadler, 2015), feeling vilified or having one's collective autonomy restricted can lead to backlash (Kachanoff et al., 2020; Peetz, Gunn, & Wilson, 2010). In this work we test whether the identity-based threats evoked by holding an equalizing interpretation lead to backlash, causing White Americans to deny historical discrimination of Black Americans and instead focus on their own victimhood (Phillips & Lowery, 2015)—the opposite aim of educators who teach about structural racism (Kendi, 2019; Tatum, 1992).

In sum, we hypothesize that having an equalizing interpretation of anti-racist messages will be associated with White Americans experiencing greater threat and backlash when seeing these messages. On the other hand, we expect that having a building-block interpretation will *not* be associated with threat and resistance—it may even be associated with message support and anti-racist action, because this interpretation resonates with the core message of anti-racism educators (Kendi, 2019).

Importantly, we conceptualize the equalizing interpretation versus building-block interpretation of anti-racist messages as two separate dimensions that people may vary independently on. Imagine two people who interpret an anti-racist message to suggest that White silence enables continued racial violence and inequity to go unchallenged (a building-block interpretation). One of these people might also interpret the message as equating silence to be as morally wrong as actually engaging in violence (an equalizing interpretation), while the other person does not. Yet another person might focus only on the idea that silence is equal in moral wrongness to violence (high equalizing interpretation) and not on the idea of silence facilitating violence (low building-block interpretation). Finally, another person might not focus on either idea at all (low in building-block and equalizing interpretations). Given this complexity, we explored whether there is an interaction between these two distinct interpretations, in terms of how White people respond to seeing an anti-racist message (versus no message). We predict that message exposure (versus no exposure) might elicit threat and denial of racism most among White Americans high in an equalizing interpretation *and* low in a building-block interpretation (Kendi, 2019).

4. Antecedents of an equalizing interpretation

We also consider what predicts whether people have a building-block vs. an equalizing interpretation. Political conservatives are often more motivated to defend the hierarchical status quo (Ho et al., 2015; Jost, Nosek, & Gosling, 2008); thus, they might be particularly sensitive to potential threats against it, leading them to interpret the intent behind messages like the PWS in more threatening ways. Similarly, White Americans high in White ethnic identification (Gunn & Wilson, 2011), collective narcissism (Marchlewska, Cichocka, Jaworska, Golec de Zavala, & Bilewicz, 2020), or anxiety about race relations (Trawalter, Richeson, & Shelton, 2009; Vorauer, 2006) might have an equalizing interpretation because they are more likely to amplify threats to their group identity. Finally, reliance on environmental heuristics (Frederick, 2005; Toplak, West, & Stanovich, 2011) may predict an equalizing interpretation, as it takes more mental steps to argue that "mild behaviors serve as a foundation for extreme behaviors" versus "mild behaviors equal extreme behaviors."

5. Current research

We had four research objectives tested across four pre-registered studies. First, we assessed whether White Americans differ in their interpretation of anti-racist messages about structural racism (i.e., in terms of their equalizing and building-block interpretations) and assessed the consequences of these interpretations for White Americans' identity threat and resistance both cross-sectionally (Study 1) and experimentally (Study 2). Second, we investigated which individual differences predict White Americans' differing interpretations (Study 1 and Study 4). Third, we experimentally tested the effect of message exposure (versus no exposure) on White Americans' level of identity threat and anti-racist attitudes and tested whether these effects are moderated by Americans' interpretation (Study 3). Fourth, we examined whether simple alterations to messages about structural racism—highlighting a building-block interpretation while differentiating between silence and violence—reduces backlash without undermining their core message or motivational effectiveness.

All studies received IRB ethics approval. All studies were pre-registered, but we made some modifications to our analysis strategies after data was collected: Please see Supplemental Analyses for point-by-point descriptions of changes made to our pre-registered analysis plans. Data, analysis code, and all materials for all studies are available on the OSF: https://osf.io/jr2t4/?view_only=8638649403294319a9d20e7097ddf123.

6. Study 1

In Study 1, we assessed how White-identified people among a representative American sample interpreted two anti-racist messages about structural racism: the Pyramid of White Supremacy and the "White silence is violence" message. We predicted that having an equalizing interpretation would be associated with White identity threat, and in turn, greater message resistance and less anti-racist motivation. We also explored antecedents of *why* White Americans might have an equalizing interpretation. We assessed interpretations of both the PWS and "White Silence" message to ensure our predicted effects were robust both when aggregating across the two messages, and when comparing effects for the relatively long PWS message and relatively short "White Silence" message separately (see Supplemental Tables 12 and 13).

6.1. Method

Participants. We used Prolific to recruit a representative and stratified sample of 647 Americans (based on age, gender, and ethnicity) between April 22nd and April 24th 2021 – this sample size was determined and collected before any analyses. We excluded 47 participants prior to analyses who failed pre-registered attention checks (https://aspredicted.org/blind.php?x=q94gm3), and/or who did not agree to release their data after reading the debriefing form. Our final sample consisted of 600 Americans ($M_{age} = 46.00$; $SD_{age} = 16.09$; 302 Female, 290 Male, 6 Non-Binary/ third gender, 2 did not disclose gender). As pre-registered, we conducted analyses using a multigroup SEM path modeling strategy which tested relations between equalizing interpretation and antecedents/outcomes for both White Americans and Americans of Color simultaneously. However, we considered analyses for Americans of Color as exploratory and had no clear a priori hypotheses for Americans of Color. Thus, we focus on the portion of results for White Americans ($N = 428$) and report the portion of results for Americans of Color ($N = 172$) in Supplementals Analyses. This sample size yielded 600 observations ensuring at least 5–10 observations per parameter for the SEM analyses we conducted with the largest number of parameters (in this case at least an $n = 420$; see Kline, 2011).

6.2. Materials

Messages about Structural Racism. Participants saw two messages about structural racism. One message was the Pyramid of White Supremacy (PWS) (see Fig. 1; sosspeace.org, 2019). The second message was "White Silence is Violence", which similarly can either be interpreted in equalizing terms (silence is literally violence) or building-block terms (by being a foundation for violence, silence is violence). Both messages were presented in random order – for each message we assessed people's (a) equalizing and building-block interpretations, (b) message resistance, and (c) the extent to which people felt motivated by the message to engage in anti-racist action.

Measures. Unless specified otherwise, all scale items were rated on a 7-point scale from 1 (strongly disagree) to 7 (strongly agree). Cronbach's alphas are reported using all responses from White Americans and Americans of Color (see Supplemental Information for Study 1 for alphas pertaining to White Americans and Americans of Color separately).

Equalizing and Building-Block Interpretation. We developed three items to assess equalizing interpretation ($\alpha_{silence} = 0.74$; $\alpha_{pws} = 0.84$) and three items to assess building-block interpretation ($_{silence} = 0.78$; $\alpha_{pws} = 0.75$; See Table 1 for all scale items). For both message contexts, pre-registered confirmatory factor analysis (Byrne, 1994a, 1994b; see Table 1) confirmed acceptable fit for a two-factor model representing equalizing interpretation and building-block interpretation when

Table 1
Standardized factor loadings derived from two independent CFAs (one for each message type) of the structural racism interpretation scale (using data from White Americans and Americans of Color; Study 1).

Context: white silence is violence	Equalizing interpretation	Building-block interpretation
1. This message equates silence as being the same as violence.	0.68	
2. This message argues that people who are silent about race issues in America are themselves engaging in racial violence.	0.82	
3. This message argues that being silent about racial issues makes you a blatant racist.	0.62	
4. This message argues that being silent about racism creates a foundation which can lead to racial violence occurring in our society.		0.8
5. This message argues that racial violence exists in a society when people are silent about racism.		0.76
6. This illustration argues that racial violence occurs in our society when it is normal for people not to confront racism when they see it.		0.66
Context: The Pyramid of White Supremacy		
1. This illustration equates indifference to prejudice with mass murder.	0.65	
2. This illustration argues that people who don't confront prejudice are White supremacists.	0.88	
3. This illustration argues that denying White privilege makes you a White supremacist.	0.88	
4. This illustration argues that showing indifference to prejudice creates a foundation which can lead to mass murder occurring in our society.		0.63
5. This illustration argues that White Supremacy exists in a society when people minimize or show indifference to prejudice.		0.77
6. This illustration argues that hate crimes occur in our society when it is normal for people not to confront prejudice when they see it.		0.77

Note. CFA Fit Indices (White Silence Context): $CFI = 0.99$, $SRMR = 0.031$, RMSEA = 0.050, $BIC = 12{,}805.12$, $\chi^2(8) = 20.12$, $p = .010$. CFA Fit Indices (PWS Context): $CFI = 0.99$, $SRMR = 0.027$, RMSEA = 0.044, $BIC = 13{,}217.491$, $\chi^2(8) = 17.20$, $p = .028$.

assessing data from White Americans and Americans of Color together (we also found evidence of structural invariance across White Americans and Americans of Color for both message types when conducting a multigroup CFA based on Ethnicity; See Supplemental Information for Study 1 for CFA details). Importantly, the two-factor model had superior fit to a one-factor model both in the White Silence” message context ($\chi^2_{\text{diff}} = 371.91, p < .001$) and the PWS message context ($\chi^2_{\text{diff}} = 407.34$, $p < .001$). Additionally, we found support for a two-factor structure when we conducted exploratory factor analysis (Carpenter, 2018) with an independent pre-registered sample of White American participants ($N = 299$; see Supplemental Study 1; https://aspredicted.org/blind.php?x=bm8fx6).

Message Resistance. We assessed message resistance with three items: “Do you agree with the message that is being sent by (message name)” (1 = strongly disagree; 6 = strongly agree; reverse coded); “Do you support the (message name) being used to teach people about prejudice?” (1 = strongly oppose; 6 = strongly support; reverse coded); “How harmful do you think the (message name) is to our society?” (1 = not at all harmful; 6 = extremely harmful; $\alpha_{\text{silence}} = 0.93$; $\alpha_{\text{pws}} = 0.94$).

Anti-Racist Motivation. We assessed how much people felt motivated by each message to engage in anti-racist actions with two items: “How much does the (message name) make you want to speak out against anti-Black racism?”; and “How much does the (message name) make you want to play an active role in challenging systemic injustices in America?” (1 = “not at all”; 6 = “very much so”; $r_{\text{silence}} = 0.91, p < .001$; $r_{PWS} = 0.88$, $p < .001$).

6.2.1. Ethnic identity threat (mediators)

Moral Identity Threat. We assessed people’s beliefs that their ethnic or racial group has been vilified as the “bad” group (adapted from Peetz et al., 2010) with 5 items. Sample item included: “Members of my ethnic/racial group are always cast as the “villains“ of our society” (α = 0.95).

Collective Autonomy Threat. We assessed collective autonomy threat with five items from Kachanoff, Taylor, Caouette, Khullar, and Wohl (2019) collective autonomy restriction scale. Sample item included: “Other groups try to control what members of my ethnic/racial group should value and believe” (α = 0.98).

6.2.2. Antecedents of an equalizing interpretation

Resistance to Environmental Heuristics. We assessed how resistant participants were to the influence of environmental heuristics using the 4-item Cognitive Reflection Test 2 (Thomson & Oppenheimer, 2016). A sample item included: If you’re running a race and you pass the person in second place, what place are you in? (intuitive wrong answer: first; correct answer: second; α = 0.59). Participants received 1 point for each correct answer (participants could have a maximum score of 4).

Collective Narcissism. We assessed collective narcissism towards one’s ethnic/racial group using five items we adapted from the short version of Golec de Zavala and colleagues’ (2009) collective narcissism scale. Sample item included: “It really makes me angry when others criticize members of my ethnic/racial group” (α = 0.92).

Ethnic Identification. We assessed identification to one’s ethnic/racial group with three items adapted from Leach et al. (2008). Sample item included: “How strongly do you identify with other members of your ethnic/racial group” (1 = not at all; 7 = very strongly; α = 0.92).

Racial Anxiety. We developed 5 face-valid items to assess anxiety about confronting race issues: Sample item included: “As a member of my ethnic/racial group, I am quite nervous about how volatile race relations are in America right now” (α = 0.80).

Political Conservatism. We assessed political orientation with three items. Two items assessed peoples’ (1) economic views, and (2) social views on a scale from 1 (*Very Liberal*) to 7 (*Very Conservative*). We also assessed general political party preference from 1 (*Strong Democrat*) to 7 (*strong Republican*; α = 0.91).

6.3. Results

Pearson correlations and descriptive statistics for White Americans are shown in Table 2 (see Supplemental Table 4 for Americans of Color).

6.3.1. Are White Americans split in how they interpret anti-racist messages?

Equalizing Interpretation. While having an equalizing interpretation was more common for the “White Silence is Violence” message ($M = 4.95$, $SD = 1.54$) than for the “Pyramid of White Supremacy” illustration ($M = 3.52$, $SD = 1.79$; F(427) = 285.35, $p <$,001), the distribution of responses approximated a normal distribution for both messages (see Fig. 2) with meaningful variation around the mean (see Fig. 2 for skewness/kurtosis statistics). Distributions also approximated a normal distribution for Americans of Color (see Supplemental Fig. 1).

Building-Block Interpretation. White Americans had a building-block interpretation of the “White Silence” message ($M = 5.83$, $SD = 1.26$), and the PWS illustration ($M = 5.55$, $SD = 1.35$) significantly above the scale midpoint for both messages (all $ps < 0.001$). The distribution of White Americans’ building-block interpretation did not reflect a normal distribution for either message. Repeated measure ANOVAs suggested that building-block ratings were significantly higher than equalization ratings for White Americans for both message types (all $ps < 0.001$).[1]

6.3.2. What predicts an equalizing interpretation?

We used a multi-group (White Americans vs. Americans of Color) SEM path-model to regress potential antecedents of equalization (i.e., resistance to environmental heuristics, collective narcissism, ethnic identification, race anxiety, and conservative ideology) onto people’s average equalizing interpretation across the two messages.[2] We controlled for people’s average building-block interpretation across the two messages.[3] The model was fully saturated, $\chi^2(0) = 0$. We focus on the portion of results pertaining to White Americans here and report the portion of results pertaining to Americans of Color in Supplemental Table 5.

Among White Americans, collective narcissism, racial anxiety, and conservative ideology significantly (positively) related to having an equalizing interpretation (see Table 3), although we note that group identification also positively related to having an equalizing interpretation in zero-order terms (See Table 2 for correlations).[4]

6.3.3. What are the consequences of an equalizing interpretation?

We used a multi-group (White Americans vs. Americans of Color) SEM path model to test whether having an equalizing interpretation was associated with (a) greater resistance to messages about structural racism, and (b) less motivation to engage in anti-racist action in response to the message. We focus on results for White Americans (see Table 4) and report results for Americans of Color in Supplemental Table 11. We formed composite scores of equalizing interpretation, building-block interpretation, message resistance, and anti-racist motivation across

[1] This ANOVA was not included in pre-registration.

[2] As pre-registered, we report supplemental antecedent analyses for both message contexts separately in Supplemental Tables 7–10. Results for White Americans were consistent across both contexts with the exception that conservatism predicted equalizing interpretation for the PWS message but not the White Silence message.

[3] Among White Americans the effects of narcissism ($p = .002$) and racial anxiety ($p < .001$) predicting equalizing interpretation remain robust *without* controlling for building-block interpretation but the positive relation with conservatism became non-significant ($p = .191$) – this additional analysis was not pre-registered.

[4] We conducted a similar multi-group SEM path analysis predicting building-block interpretation from antecedents while controlling for equalization. Narcissism, ethnic identification, and conservative ideology were all significantly negatively associated with having a building-block interpretation for White Americans (See Supplemental Table 6).

Table 2
Descriptive statistics and pearson-correlations (White Americans, Study 1).

	M	*SD*	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1. Equalizing Interpretation (PWS)	3.53	1.79	1														
2. Equalizing Interpretation (Silence)	4.95	1.54	0.46***	1													
3. Building-Block Interpretation (PWS)	5.5	1.35	0.24***	0.23***	1												
4. Building-Block Interpretation (Silence)	5.83	1.26	0.03	0.29***	0.47***	1											
5. Message Resistance (PWS)	3.41	1.81	0.44***	0.21***	−0.34***	−0.31***	1										
6. Message Resistance (Silence)	3.72	1.73	0.33***	0.19***	−0.26***	−0.32***	0.76***	1									
7. Anti-Racist Motivation (PWS)	3.49	1.77	−0.29***	−0.20***	0.25***	0.26***	−0.73***	−0.70***	1								
8. Anti-Racist Motivation (Silence)	3.23	1.85	−0.21***	−0.13**	0.19***	0.24***	−0.60***	−0.78***	0.81***	1							
9. Collective Autonomy Threat	3.15	2.02	0.45***	0.24***	−0.09	−0.18***	0.60***	0.55***	−0.44***	−0.40***	1						
10. Moral Identity Threat	3.42	1.98	0.46***	0.24***	−0.13**	−0.21***	0.63***	0.60***	−0.52***	−0.47***	0.80***	1					
11. Resistance to Environmental Heuristics	2.42	1.09	−0.09	0.05	0.10*	0.08	−0.11*	−0.04	0.03	0.003	−0.14**	−0.07	1				
12. Collective Narcissism	2.33	1.49	0.39***	0.16**	−0.26***	−0.26***	0.48***	0.41***	−0.34***	−0.26***	0.65***	0.65***	−0.10*	1			
13. Group Identification	3.98	1.73	0.27***	0.09†	−0.29***	−0.21***	0.38***	0.35***	−0.24***	−0.18***	0.44***	0.46***	−0.07	0.62***	1		
14. Racial Anxiety	3.51	1.55	0.38***	0.19**	−0.07	−0.14**	0.47***	0.43***	−0.35***	−0.31***	0.65***	0.66***	−0.13**	0.51***	0.37***	1	
15. Political Conservatism	38.33	31.09	0.32***	0.14**	−0.20***	−0.23***	0.63***	0.55***	−0.54***	−0.48***	0.61***	0.65***	−0.07	0.50***	0.41***	0.52***	1

Note. †$p < .10$, * $p < .05$, ** $p < .01$, *** $p < .001$.

the two message contexts.[5] We controlled for participants' building-block interpretation and tested models both excluding (pre-registered) and including (not pre-registered) antecedent variables as covariates (both models were both fully saturated).

White Americans who had a greater equalizing interpretation across the two messages were significantly more likely to resist these messages and were significantly less motivated by the messages to engage in anti-racist action. In contrast, holding a building-block interpretation was associated with less message resistance and greater motivation to engage in anti-racist action. The significance of these effects did not change controlling for potential antecedents of equalization. Of note, Americans of Color who tended to have an equalizing interpretation were also more resistant to messages about structural racism; see Supplemental Table 11).

6.3.4. *Does ethnic identity threat mediate the effects of equalization?*

We used a multi-group (White Americans vs. Americans of Color) SEM path model (Fig. 3) to test if having an equalizing interpretation was indirectly associated with message resistance and anti-racist motivation through ethnic identity threat (i.e., moral identity threat and collective autonomy threat entered as parallel mediators). While we had a theoretical basis for the order of variables in our model, this model was only one of several models possible because all measures were assessed cross-sectionally (Fiedler, Harris, & Schott, 2018). Thus, we cannot infer causal directionality of any paths within. We focus on results for White Americans and report exploratory results for Americans of Color in Supplemental Fig. 2. We again formed composite scores of equalizing interpretation, building-block interpretation, message resistance and anti-racist motivation across the two message contexts.[6] We also controlled for the effects of building-block interpretation on both mediators and both outcomes. We estimated indirect effects using 5000 boot-strapped samples.

As predicted, having an equalizing interpretation was positively associated with moral identity threat ($b = 0.51, p < .001$, 95% CI [0.42, 0.59]) and collective autonomy threat ($b = 0.47, p < .001$, 95% CI [0.39, 0.55]). Moral identity threat ($b = 0.34$, $p < .001$, 95% CI [0.23, 0.45]) and collective autonomy threat ($b = 0.20$, $p < .001$, 95% CI [0.09, 0.33]) were associated with greater resistance to messages about structural racism. The indirect relation between having an equalizing interpretation and message resistance through moral identity threat (*indirect effect* = 0.17, 95%CI [0.11, 0.23]) and collective autonomy threat (*indirect effect* = 0.09, 95%CI [0.04, 0.16]) were significant. The direct relation between having an equalizing interpretation and message resistance accounting for both threats was significant ($b = 0.25$, $p <$ .001, 95% CI [0.17, 0.34]).

White Americans who perceived moral identity threat were less motivated by the messages to engage in anti-racist action ($b = -0.37$, $p < .001$, 95% CI [−0.51, −0.23]), and the indirect relation between having an equalizing interpretation and motivation through moral identity threat was significant (*indirect effect* = −0.19, 95%CI [−0.27, −0.11]). Collective autonomy threat was not related to anti-racist motivation ($b = -0.05$, $p = .436$, 95% CI [−0.19, 0.08]), and the indirect relation between equalizing interpretation and motivation through collective autonomy threat was non-significant (*indirect effect* = −0.03, 95% CI [−0.09, 0.04]). The direct relation between equalizing interpretation and motivation accounting for both identity-based threats was significant ($b = -0.15$, $p = .003$, 95% CI [−0.24, −0.050]).

[5] We also report analyses for each message context separately in Supplemental Tables 12 and 13. Equalization was significantly positively related to message resistance and significantly negatively related to anti-racist motivation among White Americans across both message contexts.

[6] In Supplemental Figs. 3 and 4 we report mediation results for each specific message separately (not using a multilevel framework). Indirect effects were consistent for both messages separately.

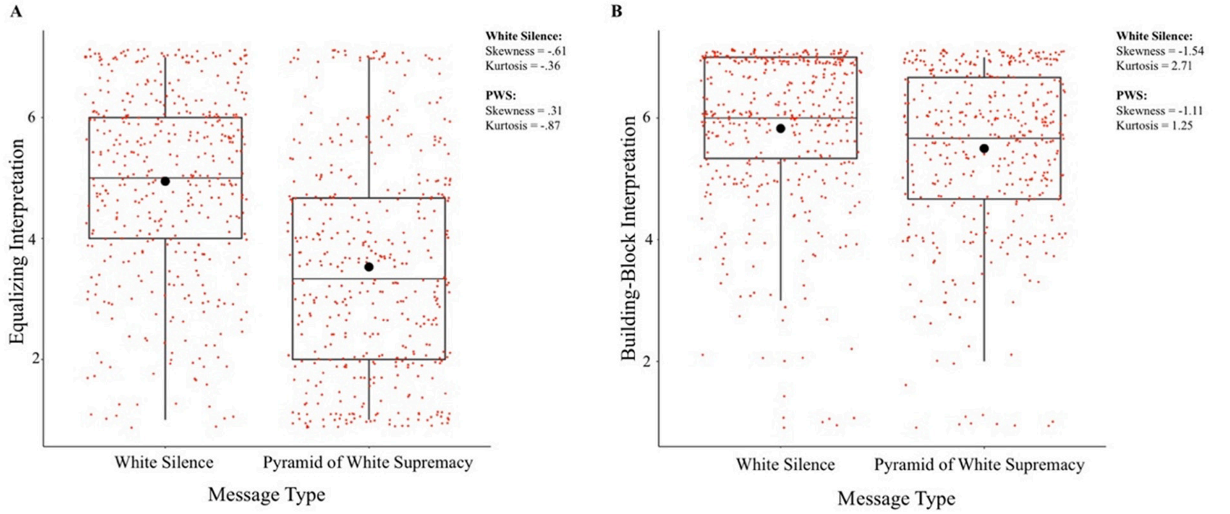


Fig. 2. White Americans' interpretation of structural racism messages (Study 1).

Table 3
Antecedents of having an equalizing interpretation (White Americans; Study 1).

	b	*p*	95% LCI	95% UCI
1. Resistance to Environmental Heuristics	−0.01	0.781	−0.09	0.07
2. Collective Narcissism	0.33	**<0.001**	0.2	0.47
3. Ethnic/Racial Identification	0.06	0.302	−0.05	0.16
4. Racial Anxiety	0.16	**0.002**	0.06	0.26
5. Political Conservatism	0.13	**0.008**	0.03	0.23
6. Building-Block Interpretation	0.4	**<0.001**	0.32	0.48

Note. Bolded values indicate a significant effect , $p < 0.05$.

7. Discussion

Study 1 supported our hypotheses: White Americans who had an equalizing interpretation of anti-racist messages about structural racism showed greater resistance to these messages and were less motivated by these messages to engage in anti-racist actions. Results also suggested that having an equalizing interpretation was indirectly related to message resistance and anti-racist motivation through White identity threat (although we cannot draw any conclusions about the causal order of these variables given the cross-sectional nature of Study 1). The effects were robust to controlling for White Americans' building-block interpretation, which was associated with greater message support and greater anti-racist motivation. Our results were also robust to controlling for factors that might lead White Americans to resist messages about structural racism: collective narcissism, ethnic identity, racial anxiety, conservatism, and general sensitivity to environmental heuristics. Among these factors, having an equalizing interpretation was most common among White Americans high in collective narcissism, racial anxiety, and conservative ideology (although the relation between conservative ideology and equalization was less robust).

8. Study 2

Study 1 provided correlational evidence that holding an equalizing interpretation of anti-racist messages about structural racism elicits ethnic identity threat among White Americans, and subsequent resistance to those messages. In Study 2 we sought experimental evidence for these effects. Using a between-subject design we compared how White Americans responded to a version of the Pyramid of White Supremacy that explicitly equated the different elements of structural racism (i.e., an *equalization condition* that we expected would evoke relatively high levels of equalizing interpretation) versus a version of the PWS that explicitly differentiated between the different elements of structural racism (i.e., a *differentiation condition* we expected would evoke relatively low levels of equalizing interpretation). We tested the effects of this manipulation on White identity threat (i.e., moral identity threat and collective autonomy threat). We also assessed different indices of message support versus backlash: this included (1) White Americans' resistance to the message itself (using items taken from Study 1), (2) White Americans' denial that Black versus White Americans face greater racial discrimination, (3) people's support for creating anti-racist spaces, (4) resistance to politically correct culture, and (5) support for collective action initiatives on behalf of White Americans. We predicted that White Americans would show greater message resistance and backlash in the equalization (vs. differentiation) condition, and that these effects would be mediated by White identity threat.

While Study 2 was pre-registered (https://aspredicted.org/blind.php?x=/VQU_FRA) we note that we pre-registered using an ANOVA framework rather than SEM framework to conduct analyses. However, we switched to the SEM framework because it allows us to test the impact of condition on all outcomes simultaneously in one analysis and is consistent with the SEM approach we preregistered in our other studies. Our results were consistent when using an ANOVA or MANOVA approach.

8.1. Method

Participants. We recruited 553 White Americans from Mechanical Turk using the CloudResearch platform between December 17th 2019 and January 2nd 2020. Our final sample consisted of 492 White Americans (M_{age} = 38.20; SD_{age} = 12.66; 296 Female, 196 Male; $N_{equivalization\ condition} = 250$; $N_{foundational\ condation} = 242$) after exclusions (see pre-registration for details). This sample ensured at least 5–10 observations per parameter for the SEM analyses we conducted with the largest number of parameters (in this case at least an $n = 350$; see Kline, 2011).

Equalizing Interpretation Manipulation. Participants were randomly assigned to view one of two different versions of the PWS. In the *equalization condition* (Fig. 4, Panel A) the pyramid included captions explicitly equating the lower levels of the pyramid (e.g., passive indifference) with the upper levels of the pyramid (e.g., racial violence). In the *differentiation condition* (Fig. 4, Panel B), the pyramid included captions explicitly stating that the different levels of the pyramid are distinct.

Table 4
Consequences of an equalizing interpretation of messages about structural racism (Study 1; White Americans).

	Excluding antecedents (pre-registered)								Including antecedents (not pre-registered)							
	Message resistance				Anti-racist motivation				Message resistance				Anti-racist motivation			
	b	*p*	95%LCI	95% UCI	*b*	*p*	95%LCI	95% UCI	*b*	*p*	95%LCI	95% UCI	*b*	*p*	95%LCI	95% UCI
1. Equalizing Interpretation	0.51	**<0.001**	0.44	0.59	−0.36	**<0.001**	−0.44	−0.27	0.31	**<0.001**	0.24	0.39	−0.22	**<0.001**	−0.31	−0.13
2. Building-Block Interpretation	−0.50	**<0.001**	−0.58	−0.43	0.37	**<0.001**	0.28	0.45	−0.34	**<0.001**	−0.41	−0.26	0.26	**<0.001**	0.17	0.35
3. Sensitivity to Environmental Heuristics									0.01	0.751	−0.06	0.08	−0.06	0.161	−0.14	0.02
4. Collective Narcissism									0.00	0.987	−0.11	0.12	0.04	0.561	−0.10	0.18
5. Ethnic/Racial Identification									0.02	0.696	−0.07	0.10	0.08	0.112	−0.02	0.19
6. Racial Anxiety									0.13	**0.002**	0.05	0.22	−0.07	0.143	−0.17	0.02
7. Political Conservatism									0.39	**<0.001**	0.31	0.47	−0.42	**<0.001**	−0.51	−0.32

Note. Bolded values indicate a significant effect , $p < 0.05$.

8.2. Measures

Manipulation checks. We used the same items and scale anchors from Study 1 to assess participants' *equalizing interpretation* (α = 0.80) and *building-block interpretation* of the PWS (α = 0.78).

Message Resistance Outcome. We used the same 3-item scale from Study 1 to assess *resistance* towards the *PWS* (α = 0.92).

Backlash versus *Support of Anti-Racism.* We assessed backlash versus support of anti-racist action in four ways: (1) White Americans denial of anti-Black versus anti-White discrimination with a one-item measure taken from Sullivan, Landau, Branscombe, & Rothschild, 2012; (2) support for creating anti-racist spaces, (3) backlash to politically correct culture, and (4) support for collective action on behalf of Whites with a scale used by Kachanoff et al. (2020). Table 5 lists all items and scale reliabilities for these measures.

White *Identity Threat (Mediators).* Moral Identity Threat (α = 0.94)[7] and collective autonomy threat (α = 0.97) were assessed as in Study 1.

8.3. Results

We tested the effect of condition (differentiation = -.5; equalization = .5) on all measures using a SEM path model. All variables were standardized prior to our analysis (see Table 6). The model was fully saturated, $\chi^2(0) = 0$. Our results also remained consistent using a univariate ANOVA based approach (see Supplemental Table 14).

Manipulation check. Our manipulation was effective: Whites Americans who were shown the equalization version of the PWS were significantly more likely to have an equalizing interpretation of the PWS than those who were shown the differentiation version. White Americans did not differ in their building-block interpretation between conditions. Of note, we did not expect differences in building-block interpretation across the two conditions given that both versions depict more extreme forms of racism at the top of the pyramid being held up by less extreme forms at the bottom.

Message Resistance and Backlash. As we predicted, White Americans who were shown the equalization version of the PWS (vs. the differentiation version) were significantly more resistant to the PWS. Importantly, beyond their resistance to the PWS itself, we observed that White Americans who were exposed to the equalization (vs. differentiation) version were also significantly less likely to acknowledge anti-Black (vs White) discrimination, were less supportive of creating anti-racist spaces, and were more likely to oppose politically correct culture. We did not observe differences between conditions in White Americans' support for collective action on behalf of White Americans (See Table 6).

Mediation Analysis. Does White Identity Threat Mediate the Effect of Equalization on Resistance? In a second SEM path-model (see Fig. 5), we tested whether the equalization (vs. differentiation) manipulation indirectly impacted people's resistance to the PWS, as well as other

[7] We also assessed moral identity threat in Study 2 using an alternative set of items that generally asks people whether they feel White Americans are regarded as immoral: e.g., "Other groups view members of my ethnic group (White Americans) as being immoral". This differs from the scale we focus on in our main analysis which assesses moral identity threat as being vilified and blamed unjustly. We focus on the latter form of moral threat given that transgressing groups in general might perceive their group to be seen as immoral – thus any description of structural racism might induce this perception in White Americans. Supporting this idea, we did not find an effect of condition on general perception of immorality ($b = 0.01$, $p = .939$, 95%CI[−0.24, 0.26]). Moreover, a general sense that one's group is perceived as immoral can actually motivate reconciliatory behaviors (Shnabel & Nadler, 2015) in contrast to vilification which can promote backlash (see Peetz et al., 2010). Consistent with this past work, we found that having a general perception that White Americans are perceived as immoral was not correlated with message resistance (r(492) = 0.06, $p = .211$), and the semi-partial correlation was actually negatively related controlling for the vilification form of moral threat (r(489) = −0.22, $p < .001$.

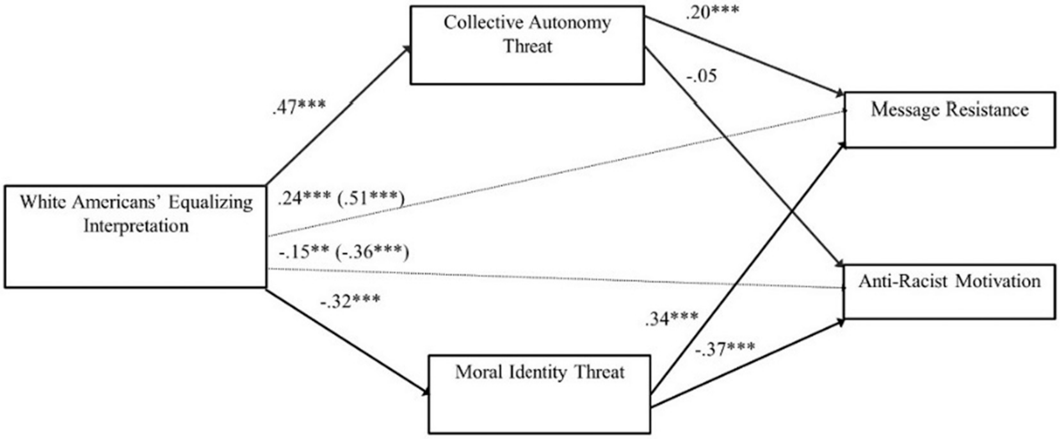


Fig. 3. The indirect relation between equalizing interpretations and resistance to messages about structural racism and anti-racist motivation through ethnic identity threat for White Americans (Study 1).

Note. Although not depicted for simplicity, the building-block interpretation was regressed onto both mediators and both outcomes. We only show the portion of results for White Americans and report the portion of results for Americans of Color in Supplemental Fig. 2. Total effects are reported in parentheses.\ $p < .10$, * $p < .05$, ** $p < .01$,*** $p < .001$.

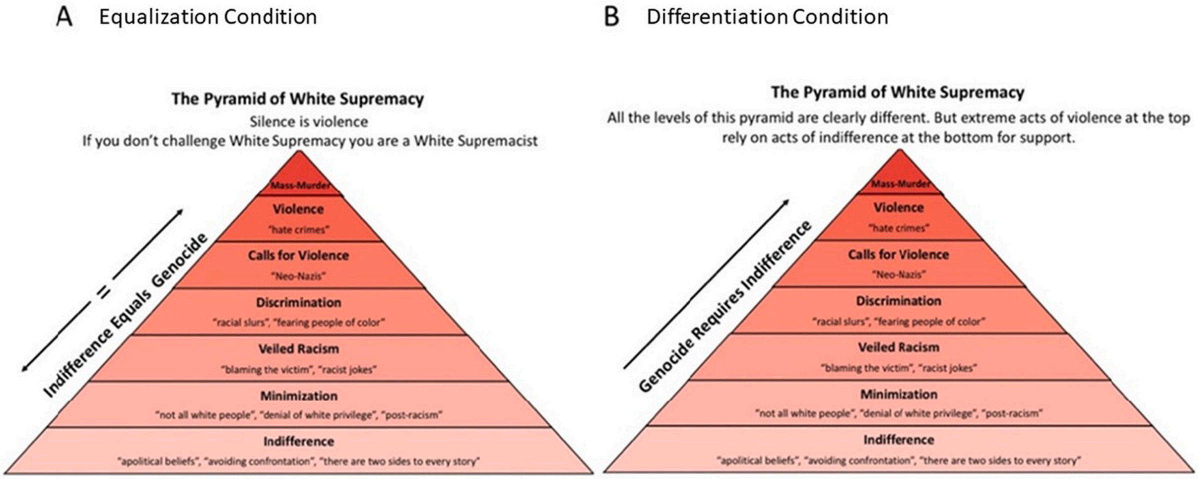

A Equalization Condition
The Pyramid of White Supremacy
Silence is violence
If you don't challenge White Supremacy you are a White Supremacist
Indifference Equals Genocide
=
Mass-Murder
Violence
"hate crimes"
Calls for Violence
"Neo-Nazis"
Discrimination
"racial slurs", "fearing people of color"
Veiled Racism
"blaming the victim", "racist jokes"
Minimization
"not all white people", "denial of white privilege", "post-racism"
Indifference
"apolitical beliefs", "avoiding confrontation", "there are two sides to every story"
B Differentiation Condition
The Pyramid of White Supremacy
All the levels of this pyramid are clearly different. But extreme acts of violence at the top rely on acts of indifference at the bottom for support.
Genocide Requires Indifference
Mass-Murder
Violence
"hate crimes"
Calls for Violence
"Neo-Nazis"
Discrimination
"racial slurs", "fearing people of color"
Veiled Racism
"blaming the victim", "racist jokes"
Minimization
"not all white people", "denial of white privilege", "post-racism"
Indifference
"apolitical beliefs", "avoiding confrontation", "there are two sides to every story"

Fig. 4. Versions of the pyramid of white supremacy used in Study 2.

backlash outcomes, by evoking two forms of White identity threat (collective autonomy threat and moral identity threat entered as parallel mediators).[8] We bootstrapped the indirect path estimates with 5000 boot-strapping samples. We covaried the two mediators, and covaried all outcomes in the model. The model was fully saturated, $\chi^2(0) = 0$. Note that we cannot infer causal order of any paths in this model (beyond the fact that message condition, randomly assigned, causally precedes all other variables).

We found that White Americans who were exposed to the equalization version of the PWS (vs. the differentiation version) felt significantly greater White collective autonomy threat but there was no significant effect of condition on moral identity threat. We found a significant indirect effect of condition on all outcomes through collective autonomy threat (but not through moral identity threat; see Table 7 for all indirect and direct condition effects).

8.4. Discussion

Study 2 builds on Study 1 by providing experimental evidence that increasing people's equalizing interpretation of an anti-racist message about structural racism (the PWS) increases their resistance to the message itself, as well as backlash against anti-racist initiatives more broadly (i.e., greater denial of anti-Black (vs. anti-White) discrimination, greater push-back to political correctness, and less support of safe spaces). We also found that evoking an equalizing interpretation increased collective autonomy threat (but not morality threat) and this mediated the effects of equalization on message resistance and backlash.

A limit of Study 2 however is its relative lack of ecological validity, in that neither version of the Pyramid of White Supremacy we showed to participants is the standard version most frequently used in the real world. Thus, in Study 3 we turned attention to experimentally testing how exposure (versus no exposure) to the PWS standard message impacts White Americans' level of White identity threat and backlash to anti-racist initiatives.

9. Study 3

Studies 1 and 2 did not directly test the effect of exposing White

[8] In Supplemental Fig. 5 we report a pre-registered mediation analysis examining the effect of condition on identity threat and backlash outcomes via changes in equalizing interpretation and building-block interpretation (entered as parallel mediators). We found significant indirect effects of condition on outcomes via equalizing interpretation only.

Table 5
Outcomes of backlash vs. support of anti-racist initiatives in Study 2.

Scale	Items
Denial of anti-Black versus *anti-White Racism (one item only)*	"Please use the scale below to fill in this statement: In society, compared with Black Americans, White Americans experience ____ discrimination? (1 = "less overall"; 4 = "As Much"; 7 = "More Overall").
Support for anti-racist spaces (α = 0.85)	1. Creating spaces in America where everyone can feel safe should be non-negotiable. 2. There should be zero-tolerance for intolerance in this country. 3. We need to publicly call out people for saying things that can be offensive. 4. We should censor language and behavior that can be disrespectful to certain groups in society.
Pushback to Politically Correct Culture (α = 0.95)	1. People in America need to grow a thicker skin. 2. People in America need to stop being overly sensitive. 3. People in America need to stop blowing little misunderstandings way out of proportion. 4. Political correctness is getting way out of hand in this country.
Support for Collective Action for Whites (Kachanoff et al., 2020s; α =0.94)	1. I think there are good reasons to have organizations that look out for the interests of Whites. 2. More needs to be done so that people remember that "White Lives" also matter. 3. Whites needs to do more to remind the world about the challenges that White people face. 4. Whites should lobby to repeal laws that give minorities an advantage on the basis that their race, at the expense of Whites.

Table 6
Descriptive statistics and main effects of condition for all measured variables by experimental condition (Study 2).

	Differentiation condition		Equalization condition		Differentiation (−0.5) vs. equalization (0.5)				
	M	*SD*	*M*	*SD*	*b*	*SE*	*p*	95% LCI	95% UCI
1. Equalizing Interpretation	3.47	1.63	4.92	1.59	0.83	0.08	**<0.001**	0.67	0.99
2. Building-Block Interpretation	5.68	1.16	5.63	1.36	−0.04	0.09	0.679	−0.21	0.14
3. Message Resistance	3.11	1.60	3.67	1.67	0.34	0.09	**<0.001**	0.16	0.51
4. Denial of anti-Black vs. Anti-White Discrimination	1.92	1.31	2.25	1.53	0.23	0.09	**0.011**	0.05	0.40
5. Support for Anti-Racist Spaces	4.73	1.53	4.42	1.64	−0.20	0.09	**0.026**	−0.38	−0.02
6. Pushback to PC Culture	4.22	1.99	4.59	1.97	0.19	0.09	**0.037**	0.01	0.36
7. Support for White Collective Action	2.70	1.77	2.89	1.75	0.11	0.09	0.244	−0.07	0.28
8. Moral Identity Threat	3.10	1.76	3.37	1.78	0.15	0.09	0.086	−0.02	0.33
9. White Collective Autonomy Threat	2.92	1.70	3.27	1.91	0.19	0.09	**0.033**	0.02	0.37

Note. Bolded values indicate a significant effect, $p < 0.05$.

Americans to anti-racist messages about structural racism (e.g., the Pyramid of White Supremacy) versus no message. Moreover, Study 2 while experimental, did not focus on the standard version of the PWS most used in the real world. Thus, in Study 3 we directly test the effect of exposure (vs. no exposure) of standard anti-racist messages on White Americans' level of White identity threat and anti-racist attitudes.

We predicted that White Americans would respond differently to these messages as a function of their interpretation. Specifically, White Americans who we found in Study 1 to show the most resistance to these messages (i.e., those high in equalizing interpretation and low in building-block interpretation) might react adversely by experiencing White identity threat and increased denial that Black Americans (versus White Americans) face greater discrimination in the United States (Phillips & Lowery, 2015). It is less clear how individuals high in equalizing interpretation and high in building-block interpretation would respond: While having an equalizing interpretation might also lead to backlash among these individuals, having a building-block interpretation might mitigate some of this backlash. Given this complexity, we pre-registered testing a three-way interaction, crossing equalizing interpretation with building-block interpretation and with message exposure. In line with the logic above, we specifically considered the simple effect test of how participants low in building-block interpretation and high in equalization to be our primary apriori contrast of interest (and the other simple effect contrasts exploratory).

9.1. Method

Study 3 was a pre-registered study with a large sample size (N = 1500 prior to pre-registered exclusions; https://aspredicted.org/blind.php?x=4jx7p5) – we recruited a large sample because interactions with categorical variables are often underpowered in social psychology (Blake & Gangestad, 2020; Ginersorolla, 2018; Simonsohn, 2014). We ensured 95% power for detecting a small effect of message exposure (vs. no exposure) on all outcomes (i.e., Cohen's d = 0.18). In supplemental analyses, we also report results of two very similar studies, as well as results from a mega-analysis (Costafreda, 2009; Curran & Hussong, 2009) using the merged data from all three studies (see Supplemental Tables 16–21 and Supplemental Fig. 8). We exclude these two supplemental studies from our primary analysis because in these earlier studies we did not pre-register our prediction of the three-way interaction and our specific focus on individuals low in building-block interpretation and high in equalizing interpretation.

Participants. We sought to recruit 1500 participants who identified as White American: In total 1594 participants accessed our survey via Prolific during the month of May 2021. Prior to analyses, we excluded participants who accessed but did not actually participate in the study, did not pass our preregistered inclusion criteria, and/or did not release their data. After exclusions 1337 White American participants were included in our final sample (M_{age} = 37.20, SD = 12.71; 631 Male, 706 Female).

Procedure. We experimentally manipulated whether we showed the PWS (Fig. 1) to participants before versus after assessing White identity threat and backlash (i.e., the message-exposure condition vs. the no-message control condition). As an index of backlash, participants completed the same denial of anti-black (versus anti-White) racism used in Study 2 (Sullivan et al., 2012). We assessed moral identity threat (α =

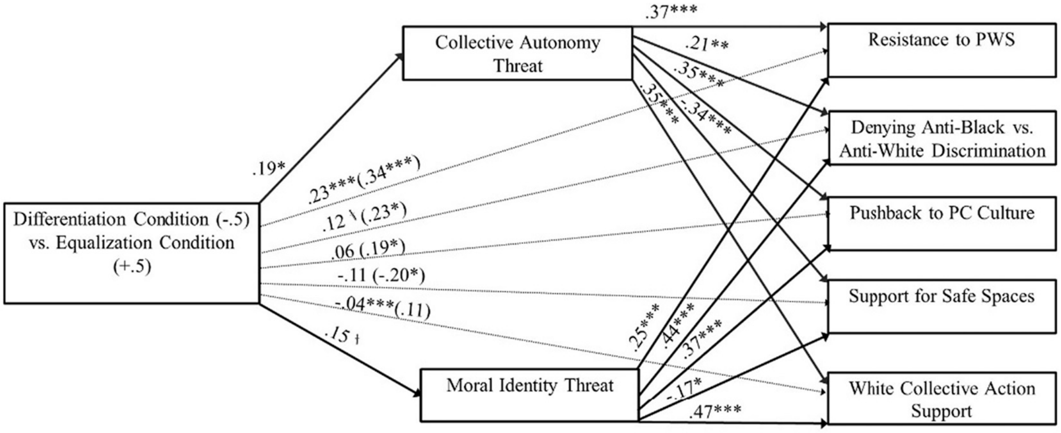


Fig. 5. Path model depicting the indirect effects of the equalization (vs. differentiation) manipulation on outcomes through white identity threat (Study 2). *Note.* Although we did not draw all covariances for visual simplicity, the two mediators covaried in the model and all five outcomes covaried in the model. Total effect reported in parenthesis. † <0.10, * < 0.05, ** < 0.01, *** < 0.001.

Table 7
Indirect effects of the equalization (vs. differentiation) condition on all mediator and outcome variables (Study 2).

	Indirect effect through moral identity threat			Indirect effect through collective autonomy threat			Direct effect of condition		
	b	95%LCI	95% UCI	*b*	95%LCI	95% UCI	*b*	95%LCI	95% UCI
Resistance to the PWS	0.04	−0.01	0.10	**0.07**	0.01	0.14	**0.23**	0.09	0.37
Denial of Anti-Black vs. Anti-White Discrimination	0.07	−0.01	0.16	**0.04**	0.003	0.09	0.12	−0.02	0.26
Support for Anti-Racist Spaces	−0.03	−0.07	0.004	**−0.07**	−0.14	−0.01	−0.11	−0.26	0.05
Pushback to PC Culture	0.06	−0.01	0.14	**0.07**	0.005	0.14	0.06	−0.07	0.19
Support for Collective Action on Behalf of Whites	0.07	−0.01	0.17	**0.07**	0.005	0.14	−0.04	−0.15	0.08

Note. Bold coefficients indicate statistical significance (0 not in the 95% confidence intervals).

0.94) and collective autonomy threat (α = 0.97; i.e., our mediators) with the measures used in Study 1.[9]

We then presented the PWS to all participants at the end of the study (i.e., participants in the message-exposure condition were re-shown the PWS a second time, while participants in the no-message control saw the PWS for the first time) and assessed participants' equalizing (α = 0.82) and building-block (α = 0.76) interpretation as in Study 1. To ensure the amount of PWS exposure was identical across conditions, participants in the no-message control were first shown the PWS on a single page (identical to what was shown to those in the message-exposure condition) and were then re-presented the PWS again on a separate page when we assessed interpretation. We also assessed White identification and political conservatism at the beginning of the survey. There were no spontaneous between-condition effects (all *ps* > 0.39). Thus, following our pre-registration, we did not control for these variables since there were no condition differences.

9.2. Results

Descriptive statistics and Pearson-Correlations are reported in Supplemental Table 15.

9.2.1. Prevalence of equalizing interpretation and building-block interpretation

Consistent with Study 1, equalizing interpretation of the PWS was about half a point below the midpoint (4) of the scale regardless of whether participants were shown the PWS before ($M = 3.51$, $SD = 1.67$) or after ($M = 3.34$ $SD = 1.68$) rating the outcome variables. There were no significant differences between conditions, $F(1,1335) = 3.60$, $p = .058$, Cohen's d = 0.10 (despite having 95% power to detect differences as small as a Cohen's d of 0.18), making moderation analysis possible (See Fig. 6). Also consistent with Study 1, people's building-block interpretation was above the midpoint (4) of the scale regardless of whether participants were shown the PWS before ($M = 5.55$, $SD = 1.22$) or after ($M = 5.59$, $SD = 1.20$) rating the outcome variables. Again, there were no significant differences between conditions, $F(1,1335) = 0.38$, $p = .539$, Cohen's d = 0.03.

9.2.2. Effect of message exposure on identity threat and anti-racist attitudes

Overall effects of condition. Exposure to the PWS message (vs. no message control) did not significantly impact collective autonomy threat ($b = 0.01$, 95% CI[−0.19, 0.19], $p = .956$), moral identity threat ($b = -0.16$, 95% CI[−0.36, 0.03], $p = .099$), or denial of anti-Black (vs. anti-White) discrimination ($b = 0.01$, 95% CI[−0.14, 0.16], $p = .907$).[10]

Overall effects of equalization and building-block interpretation. We tested the relation between equalizing interpretation and building-block interpretation and outcomes (accounting for condition). Regardless of whether outcomes were assessed before or after people first saw the PWS, having an equalizing interpretation was positively associated with

[9] In Study 3 we pre-registered an exploratory outcome– White American's belief that White Americans should speak out against racism. A sample item included "White Americans should do all they can to speak up against racial inequities". While equalization was significantly negatively associated with White Americans' support of speaking against racism, we found no significant effect of message exposure condition on this outcome, nor were there any significant condition by equalization by building-block interpretation interactions (see Supplemental Table 22 for detailed results).

[10] We only pre-registered the three-way interaction analysis, and made no pre-registered predictions regarding main effects. Still, it is informative to report the main effects of condition, equalization interpretation, and building-block interpretation on outcomes prior to including the interaction in the model.

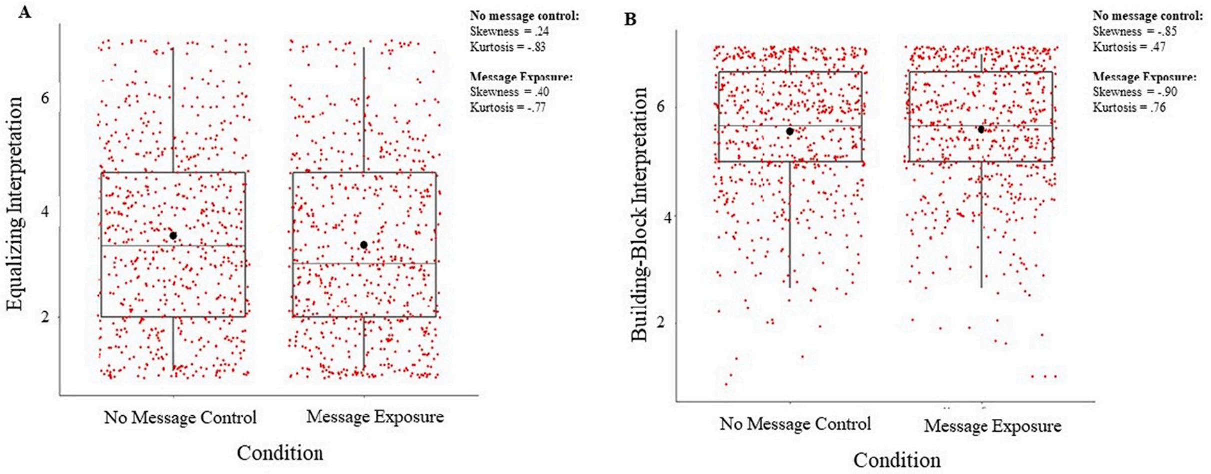


Fig. 6. White Americans' equalizing interpretation and building-block interpretation of the pyramid of white supremacy (Study 3).

Table 8
White identity threat and denial of anti-black vs. anti-white discrimination as a function of the three-way condition by building-block interpretation by equalizing interpretation interaction (Study 3).

	Collective autonomy restriction threat	Moral identity threat	Denial of anti-black vs. anti-white discrimination
	b (95% CI)	*b* (95% CI)	*b* (95% CI)
Condition	0.14 (−0.03, 0.30)	−0.03 (−0.20, 0.14)	0.09 (−0.04, 0.23)
Building-Block Interpretation	−0.34 (−0.41, −0.27)***	−0.29 (−0.36, −0.21)***	−0.31 (−0.36, −0.25)***
Equalizing Interpretation	0.52 (0.47, 0.58)***	0.52 (0.47, 0.57)***	0.38 (0.34, 0.43)***
Condition X Building-Block Interaction	−0.04 (−0.18, 0.10)	−0.08 (−0.22, 0.07)	−0.05 (−0.16, 0.06)
Condition X Equalization Interaction	0.04 (−0.06, 0.14)	−0.001 (−0.11, 0.10)	0.01 (−0.07, 0.09)
Building-Block X Equalization Interaction	−0.01 (−0.05, 0.03)	0.01 (−0.03, 0.06)	−0.03 (−0.07, −0.0005)*
3-Way Interaction	−0.10 (−0.18, −0.01)*	−0.09 (−0.18, −0.01)*	−0.03 (−0.10, 0.03)

Note.\p < .10, * *p* < .05, ** *p* < .01, *** *p* < .001.

collective autonomy threat (*b* = 0.52, 95% CI[0.47, 0.57], *p* < .001), moral identity threat (*b* = 0.52, 95% CI[0.47, 0.57], *p* < .001), and denying that Black Americans face greater discrimination than White Americans (*b* = 0.37, 95% CI[0.33, 0.41], *p* < .001). In contrast, having a building-block interpretation was negatively associated with collective autonomy threat (*b* = −0.34, 95% CI[−0.41, −0.27], *p* < .001), moral identity threat (*b* = −0.29, 95% CI[−0.36, −0.22], *p* < .001), and denying anti-Black discrimination (*b* = −0.29, 95% CI[−0.35, −0.23], *p* < .001).

Interaction Model. We tested the three-way interaction effect between condition (effect coded such that the no-message control was the reference group), building-block interpretation, and equalizing interpretation (See Table 8 for interaction model and Table 9 for simple effects).

The three-way interaction between message exposure, building-block interpretation, and equalization was significant when predicting collective autonomy threat (*b* = −0.10, 95% CI[−0.18, −0.01], *p* = .022) and significant when predicting moral identity threat (*b* = −0.09, 95% CI[−0.18, −0.01], *p* = .034). The three-way interaction was non-significant when predicting denial of anti-Black discrimination (*b* = −0.03, 95% CI[−0.10, 0.03], *p* = .340).

We probed the significant three-way interactions of equalizing interpretation, building-block interpretation and message exposure on collective autonomy threat and morality threat. As predicted, White Americans high in equalizing interpretation and low in building-block interpretation felt significantly greater collective autonomy threat (*b* = 0.44, 95% CI[0.03, 0.85], *p* = .033) if they were exposed to the PWS. No other simple effects were significant (ps >0.17). Counter to prediction we found no significant effect of message exposure on moral identity threat (*b* = 0.25, 95% CI[−0.17, 0.67], *p* = .249) for individuals low in building-block interpretation and high in equalization. However, we did find that message exposure actually reduced moral identity threat in White Americans high in building-block interpretation and high in equalization (*b* = −0.32, 95% CI[−0.63, −0.006], *p* = .046). No other simple effects were significant (ps > 449).

*Mediation Analysis.*Using PROCESS (model 12; Hayes, 2017) and 5000 boot-strapping samples, we tested if message exposure was indirectly associated with denying anti-Black (versus anti-White) discrimination through increased moral identity threat and/or collective autonomy threat (entered as parallel mediators)[11] among White Americans low in building-block interpretation and high in equalizing interpretation. We allowed the interaction between building-block interpretation and equalizing interpretation to moderate the relation between condition and both types of White identity threat (i.e., the a1 and a2 paths) and the relation between condition and denying anti-Black (vs. anti-White) discrimination (i.e., the c' path). In this model, we can infer some causal ordering since we experimentally manipulated experimental condition. We test whether condition (i.e., assessing ethnic identity threat and discrimination denial before or after message exposure) influences White identity threat and denial of anti-Black

[11] As we pre-registered we also conducted a supplemental moderated moderated mediation analysis in which we used a composite score of White identity Threat by taking the mean of collective autonomy threat and moral identity threat. All thought the IMMM was significant (−0.04, 95%CI[−0.08, −0.003]) the simple indirect effect of condition on denial for White Americans low in equalization and high in building-block interpretation was non-significant (0.15, 95%[−0.02, 0.32]); see Supplemental Fig. 7).

Table 9
Means and simple effects of no-message-control vs. message-exposure as a function of building-block interpretation and equalizing interpretation (Study 3).

Simple effects as a function of message interpretation	Low building-block interpretation (−1sd)		High building block interpretation (+1sd)	
	Low EQ (−1sd)	High EQ (+1sd)	Low EQ (−1sd)	High EQ (+1sd)
	b (95% CI)	*b* (95% CI)	*b* (95% CI)	*b* (95% CI)
Collective Autonomy Threat	−0.08 (−0.39, 0.24)	0.44* (0.04, 0.85)	0.22 (−0.09, 0.54)	−0.05 (−0.35, 0.26)
Moral Identity Threat	−0.12 (−0.45, 0.20)	0.25 (−0.17, 0.67)	0.06 (−0.27, 0.38)	−0.32* (−0.63, −0.01)

Means by condition and message interpretation	Low building-block interpretation (−1sd)		High building block interpretation (+1sd)	
–	Low EQ (−1sd)	High EQ (+1sd)	Low EQ (−1sd)	High EQ (+1sd)
	Estimated mean	Estimated mean	Estimated mean	Estimated mean
Collective Autonomy Threat				
Control	2.39	**3.94**	1.47	3.31
Message	2.32	**4.38**	1.69	3.26
Moral Identity Threat				
Control	2.82	4.32	1.99	**3.97**
Message	2.7	4.57	2.04	**3.65**

Note. Bolded values indicate a significant difference in estimated marginal means of the simple effect contrast ($p < .05$).

Table 10
Indirect effects through collective autonomy threat and moral identity threat as a function of equalization and building-block interpretation (Study 3).

Mediators	Low building-block interpretation (−1sd)		High building block interpretation (+1sd)	
	Low EQ (−1sd)	High EQ (+1sd)	Low EQ (−1sd)	High EQ (+1sd)
Indirect Effect Via Collective Autonomy Threat	−0.02 [−0.11, 0.07]	**0.11 [0.01, 0.24]**	0.06 [−0.01, 0.13]	−0.01 [−0.12, 0.09]
Indirect Effect Via Moral Identity Threat	−0.02[−0.09, 0.04]	0.04 [−0.03, 0.13]	0.01 [−0.04, 0.07]	−0.06 [−0.13, 0.01]

Note. Bolding indicates a significant indirect effect (i.e., 0 not within the 95% confidence interval).

discrimination. In this design it is not possible for ethnic threat or discrimination denial to influence condition assignment. However, because White Identity threat and denial of anti-Black discrimination are assessed as the same time (before or after message exposure) we cannot infer the proposed causal ordering in which White identity threat precedes denial of anti-Black discrimination. We based this ordering on research showing that White identity threat can lead to defensive responding in White Americans (e.g., Kachanoff et al., 2020; Peetz et al., 2010; Unzueta & Lowery, 2008), but we acknowledge that an alternative model (in which discrimination denial precedes ethnic identity threat) cannot be ruled out from this experimental design (Fiedler et al., 2018).

Although a small effect with a wide 95% confidence interval, the index of moderated moderated mediation (IMMM) did not contain zero for the path through collective autonomy restriction (IMMM = -0.02, 95% CI [-0.05, -0.0003]) but did contain zero for the path through moral identity threat (IMMM = -0.02, 95% CI [-0.04, 0.0000]). This suggested significant moderated moderated mediation via collective autonomy threat, and supported us probing indirect effects for White Americans as a function of their equalizing and building-block interpretations (Hayes, 2017). As we predicted, for White Americans high in equalizing interpretation and low in building-block interpretation, there was a significant indirect effect of message exposure on denying anti-Black (versus anti-White) discrimination through collective autonomy threat but not moral identity threat (See Table 10 for detailed results). There were no significant indirect effects through either mediator for people low in building-block interpretation and low in equalizing interpretation; high in building-block interpretation and high in equalizing interpretation; or high in building block interpretation and low in equalizing interpretation (Fig. 7).

10. Discussion

Overall, the results of Study 3 suggest that the effect of brief exposure to anti-racist messages about structural racism on White Americans' levels of White identity threat and willingness to acknowledge versus deny anti-black racism is small. We found no overall main effects of condition in the study for any outcome. This said, we do find some evidence that White Americans may react differently to messages about structural racism as a function of how they interpret those messages. White Americans who had a relatively high equalizing interpretation (and a relatively low building-block interpretation) experienced the threat of losing their freedom to express White identity when shown these messages, and this in turn, was associated with increased denial of anti-black racism. People relatively high in equalizing interpretation and relatively low in building-block interpretation made up a small portion of respondents. Specifically, 1.2% of the sample were 1SD above the sample mean in equalizing interpretation and 1SD below the sample mean in building-block interpretation, and similarly, 3.4% of the sample were equal to or above the scale midpoint on equalizing interpretation and less than the scale midpoint on building-block interpretation.[12] Still for this relatively small sub-set of White Americans, exposure to anti-racist messages appears to make them more resistant to anti-racist ideas and initiatives.

It is notable that we did not find that message exposure caused White Americans high in equalizing interpretation and low in building-block

[12] 4.4% of the sample were 1 SD below the sample mean in equalizing interpretation and 1SD below the sample mean in building-block interpretation; 5.5% of the sample were 1 SD above the sample mean in building-block interpretation and 1 SD below the sample mean in equalizing interpretation; 5.3% of the sample were 1 SD above the sample mean in building-block interpretation and 1SD above the sample mean in equalizing interpretation. In terms of the scale midpoint, 9.7% of the sample were below the scale midpoint in equalizing interpretation, and below the scale midpoint in building-block interpretation; 50% of the sample were above or equal to the scale midpoint in building-block interpretation and below the scale midpoint in equalizing interpretation; finally, 36.9% of the sample were above or equal to the scale midpoint in building-block interpretation and above or equal to the scale midpoint in equalizing interpretation.

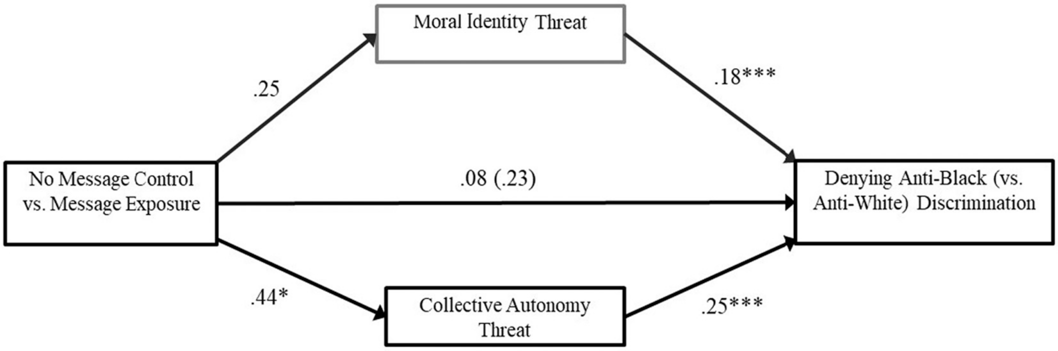

Moral Identity Threat
.25
.18***
No Message Control vs. Message Exposure
.08 (.23)
Denying Anti-Black (vs. Anti-White) Discrimination
.44*
Collective Autonomy Threat
.25***

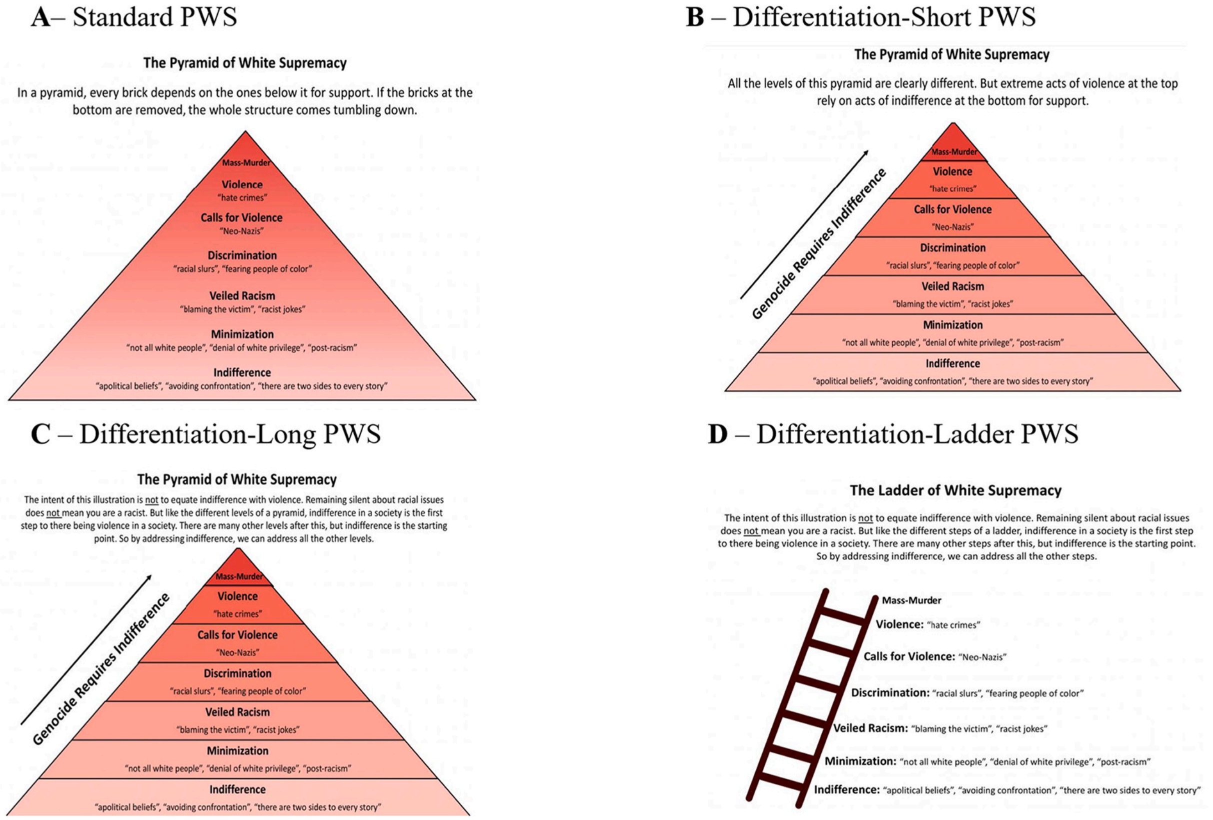
Fig. 7. Indirect effect of message exposure on denying anti-black vs. anti-white discrimination for White Americans low in building-block interpretation (−1sd) and high in equalizing interpretation (+1sd; Study 3).
Note. Total effect reported in parenthesis. †$p < .10$, * $p < .05$, ** $p < .01$, *** $p < .001$.

interpretation to experience greater moral identity threat (seeing White Americans as vilified). Moreover, White Americans high in equalizing interpretation and high in building-block interpretation reported *less* moral threat if exposed to the message. Thus, being high in building-block interpretation might override any reactance that White Americans experience from holding an equalizing interpretation.

11. Study 4

Study 3 revealed that brief exposure to anti-racist messages about structural racism have few negative (or positive) short-term effects on White Americans in general but might elicit some backlash among White Americans high in equalizing interpretation and low in building-block interpretation. Thus, in Study 4 we tested whether we could make small modifications to anti-racist messages to amplify their effectiveness in motivating anti-racist action and support. Specifically, in Study 4a and Study 4b we tested whether modifying messages about structural racism to still emphasize the structural relation between indifference and racial violence (reinforcing a building-block interpretation) while explicitly acknowledging that the different dimensions of a racist society are distinct (minimizing an equalizing interpretation) could decrease message resistance while increasing message effectiveness in promoting anti-racist motivation.

We tested this approach with the PWS (Study 4a), and the “White Silence is White Violence” message (Study 4b). In both studies we used a within-subjects design such that participants rated their interpretation and reaction to all message variants: we selected this approach to (1) maximize statistical power to detect potential differences between


A– Standard PWS
The Pyramid of White Supremacy
Mass-Murder
Violence
“hate crimes”
Calls for Violence
“Neo-Nazis”
Discrimination
“racial slurs”, “fearing people of color”
Veiled Racism
“blaming the victim”, “racist jokes”
Minimization
“not all white people”, “denial of white privilege”, “post-racism”
Indifference
“apolitical beliefs”, “avoiding confrontation”, “there are two sides to every story”
B – Differentiation-Short PWS
The Pyramid of White Supremacy
Genocide Requires Indifference
Mass-Murder
Violence
“hate crimes”
Calls for Violence
“Neo-Nazis”
Discrimination
“racial slurs”, “fearing people of color”
Veiled Racism
“blaming the victim”, “racist jokes”
Minimization
“not all white people”, “denial of white privilege”, “post-racism”
Indifference
“apolitical beliefs”, “avoiding confrontation”, “there are two sides to every story”
C – Differentiation-Long PWS
The Pyramid of White Supremacy
Genocide Requires Indifference
Mass-Murder
Violence
“hate crimes”
Calls for Violence
“Neo-Nazis”
Discrimination
“racial slurs”, “fearing people of color”
Veiled Racism
“blaming the victim”, “racist jokes”
Minimization
“not all white people”, “denial of white privilege”, “post-racism”
Indifference
“apolitical beliefs”, “avoiding confrontation”, “there are two sides to every story”
D – Differentiation-Ladder PWS
The Ladder of White Supremacy
Mass-Murder
Violence: “hate crimes”
Calls for Violence: “Neo-Nazis”
Discrimination: “racial slurs”, “fearing people of color”
Veiled Racism: “blaming the victim”, “racist jokes”
Minimization: “not all white people”, “denial of white privilege”, “post-racism”
Indifference: “apolitical beliefs”, “avoiding confrontation”, “there are two sides to every story”

In a pyramid, every brick depends on the ones below it for support. If the bricks at the bottom are removed, the whole structure comes tumbling down.

All the levels of this pyramid are clearly different. But extreme acts of violence at the top rely on acts of indifference at the bottom for support.

The intent of this illustration is not to equate indifference with violence. Remaining silent about racial issues does not mean you are a racist. But like the different levels of a pyramid, indifference in a society is the first step to there being violence in a society. There are many other levels after this, but indifference is the starting point. So by addressing indifference, we can address all the other levels.

The intent of this illustration is not to equate indifference with violence. Remaining silent about racial issues does not mean you are a racist. But like the different steps of a ladder, indifference in a society is the first step to there being violence in a society. There are many other steps after this, but indifference is the starting point. So by addressing indifference, we can address all the other steps.

Fig. 8. Versions of the pyramid of white supremacy used in Study 4a.

variants; (2) to test whether within individuals, people responded most favorably to messages they interpreted as least equivalizing; and (3) because of recent work suggesting that repeated measure designs yield the same results as between-subject designs (i.e., with minimal consistency pressures) but greater precision (Clifford, Sheagley, & Piston, 2021).

11.1. Study 4a

11.1.1. Method

Procedure. We showed participants the standard PWS (Fig. 8, panel A), and three new variants that explicitly differentiated passive indifference from active racism while still emphasizing their structural relation (see Fig. 8, Panels B—D). We explored different variants in the hope that at least one version would effectively reduce people's equalizing interpretation relative to the standard PWS. The "*differentiation-short*" variant (Panel B) –emphasized the different levels of the PWS being distinct by drawing lines between each level and stating that "each level is clearly different". This PWS variant still emphasized the building-block interpretation by showing an upward arrow of bottom levels supporting the upper levels, and stating that the upper levels require the bottom levels for support. The "*differentiation-long*" variant (Panel C) was similar but also explicitly stated that being indifferent does not make someone an active racist in terms of their moral character. Finally, the "*differentiation-ladder"* variant (Panel D) had the same caption as the differentiation-long variant but used a ladder rather than a pyramid. We were curious to see if a ladder metaphor might effectively illustrate how one aspect of a White Supremacist system leads to another while using physically 'detached' rungs to emphasize differentiation.

Participants. We recruited 451 White Americans from Mechanical Turk using the CloudResearch platform completed the study between September 2nd – 3rd 2020. We determined sample size and concluded data collection prior to analyses. After pre-registered exclusions conducted prior to analyses (https://aspredicted.org/blind.php?x=ek3ki8), our final sample consisted of 419 White Americans ($M_{age} = 43.15$; $SD_{age} = 13.06$; 229 Female, 190 Male). This sample size yielded 1676 observations (because of repeated-measures design) ensuring at least 5–10 observations per parameter for our planned SEM analyses (in this case at least an $n = 120$; see Kline, 2011).

Measures. For each variant, we assessed participants' equalizing ($\alpha_{standard} = 0.85$; $\alpha_{short} = 0.84$; $\alpha_{long} = 0.84$; $\alpha_{ladder} = 0.84$) and building-block interpretation ($\alpha_{standard} = 0.81$; $\alpha_{short} = 0.80$; $\alpha_{long} = 0.81$; $\alpha_{ladder} = 0.78$).[13] We assessed message resistance with one item: "Do you support this illustration being used as an anti-racism teaching tool in colleges, universities, and organizations?" (rated (1) "strongly oppose" to (6) "strongly support"). We assessed how much each message motivated people to want to engage in anti-racist action with the two items used in Study 1 ($r_{standard} = 0.91$; $r_{\_short} = 0.93$; $r_{\_long} = 0.92$; $r_{ladder} = 0.92$). We also ensured that our modifications did not undermine the extent to which the message effectively communicated the harm of passive indifference with one item: "To what extent do you think this illustration conveys the harm of remaining silent about race related social issues?" (rated from (1) "not at all" to (6) "very much so).

Antecedents of Equalizing Interpretation. Similar to Study 1, we assessed potential antecedents of equalizing interpretation: resistance to environmental heuristics, collective narcissism, ethnic identification, and political conservativism.[14] Replicating Study 1, collective narcissism and political conservatism were positively associated with having an equalizing interpretation across the different variants both with and without controlling for building-block interpretation. Resistance to environmental heuristics and ethnic identification were not associated with equalization (see Supplemental Tables 25–27, for reliabilities and detailed results).

11.1.2. Results

Descriptive statistics and Pearson-Correlations are summarized in Supplemental Tables 23–24. Within a multi-level SEM path model accounting for repeated observations nested within person, we tested the effect of each PWS variant (represented as 3 dummy variables contrasting the standard version to each variant) on all measures (see Study 4a supplemental information for ICCs of nested variables and details about the MLM model). Mean-ratings for all outcomes by PWS variant are summarized in Table 11.[15]

11.1.2.1. Is it possible to reduce equalizing interpretations while maintaining building-block interpretations?. Equalizing Interpretation. The differentiation-short PWS did not significantly reduce having an equalizing interpretation relative to the standard version ($b = 0.01$, $p = .868$, 95% CI [−0.06, 0.07]). However, the differentiation-long PWS ($b = -0.22$, $p < .001$, 95% CI [−0.29, −0.15]), and the differentiation ladder PWS ($b = -0.30$, $p < .001$, 95% CI [−0.37, −0.21]) both elicited significantly less equalizing interpretation relative to the standard PWS.

Building-Block Interpretation. The differentiation-short PWS was significantly more effective than the standard version in eliciting a building-block interpretation ($b = 0.15$, $p < .001$, 95% CI [0.08, 0.22]). Importantly, the differentiation-long PWS—which was effective in reducing equalizing interpretation—did *not* reduce having a building-block interpretation ($b = -0.01$, $p = .745$, 95% CI [−0.08, 0.06]). However, the differentiation-ladder PWS was less effective than the standard version in eliciting a building-block interpretation ($b = -0.25$, $p < .001$, 95% CI [−0.34, −0.17]).

Harm of Indifference. Compared to the standard PWS, participants found the differentiation-short PWS ($b = 0.11$, $p = .001$, 95% CI [0.05, 0.17]) and the differentiation-long PWS ($b = 0.15$, $p < .001$, 95% CI [0.08, 0.22]) to be significantly more effective at communicating the harm of indifference. However, the differentiation-ladder PWS was significantly less effective at communicating the harm of indifference relative to the standard PWS ($b = -0.09$, $p = .040$, 95% CI [−0.17, −0.004]).

Key Outcomes. Compared to the standard PWS, only the differentiation-long PWS significantly reduced message resistance ($b = -0.12$, $p < .001$, 95% CI [−0.18, −0.06]) and motivated greater anti-racist action intentions ($b = 0.09$, $p < .001$, 95% CI [0.04, 0.13].

[13] In our pre-registration we initially planned to use a difference score between equalization and building-block interpretation (i.e., the relative tendency for people to hold an equalization versus building-block interpretation) in our main text analyses. However, relative difference scores have limitations: although someone rating a 6 to both measures likely reflects a different psychology to someone rating a one to both measures, they are treated as equivalent with a difference score. Thus, we report analyses using the difference score in Supplemental Analyses (see Supplemental Tables 25, 28, 30, and 31 and Supplemental Fig. 9) and treat equalization and building-block interpretation as separate variables in our main text analysis. We pre-registered that we would conduct a CFA on participants' responses to the standard PWS to ensure a 2-factor model representing equalization and building-block interpretation would fit the data well: the CFA model had acceptable fit ($CFI = 0.99$, $SRMR = 0.029$, $RMSEA = 0.058$, $BIC = 8968.50$, $\chi^2(8) = 19.32$, $p = .013$).

[14] We note that in the pre-registration of Study 4a we did not specify that conservative ideology would be treated as an antecedent, but we include it as one since it was pre-registered as an antecedent in Study 1. The relations with the other antecedent variables remain consistent when not accounting for conservative ideology.

[15] We note that results remain consistent when analyzing each outcome separately with independent multilevel regression models or when using a repeated measures ANOVA to test differences between variants. Our results also remain consistent controlling for antecedents with the exception that the effect of the differentiation-long variant (vs. standard PWS variant) on anti-racist motivation becomes non-significant but remains trending in the predicted direction (See Supplemental Table 30).

Table 11
Descriptive statistics for all measured variables by message type (Study 4a).

	Standard PWS		Differentiation-short PWS		Differentiation-long PWS		Differentiation-Ladder	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
1. Equalizing Interpretation	3.54^{a}	1.74	3.55^{a}	1.74	3.16^{b}	1.68	3.02^{c}	1.66
2.Building-Block Interpretation	5.34^{a}	1.34	5.53^{b}	1.25	5.32^{a}	1.30	5.00^{c}	1.36
3. Harm of Indifference	4.08^{a}	1.54	4.25^{b}	1.57	4.32^{b}	1.54	3.95^{c}	1.54
4. Message Resistance	3.43^{a}	1.68	3.36^{a}	1.75	3.23^{b}	1.71	3.42^{a}	1.62
5. Anti-Racist Motivation	3.54^{a}	1.72	3.55^{a}	1.74	3.68^{b}	1.74	3.46^{a}	1.67

Note. Within a row, means with different subscripts are significantly different ($p < .05$).

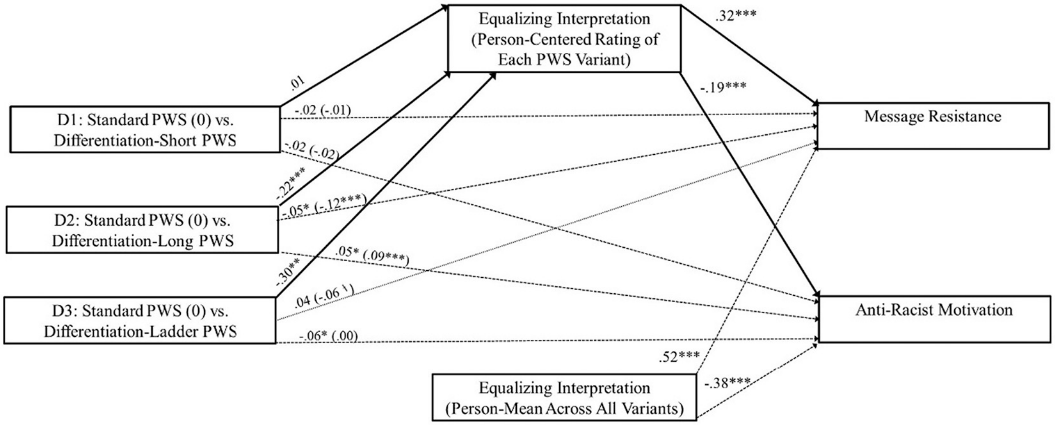

Equalizing Interpretation (Person-Centered Rating of Each PWS Variant)
.32***
-.19***
.01
D1: Standard PWS (0) vs. Differentiation-Short PWS
-.02 (-.01)
-.02 (-.02)
Message Resistance
-.22***
-.05* (-.12***)
D2: Standard PWS (0) vs. Differentiation-Long PWS
.05* (.09***)
-.30**
.04 (-.06 \)
D3: Standard PWS (0) vs. Differentiation-Ladder PWS
-.06* (.00)
Anti-Racist Motivation
.52***
-.38***
Equalizing Interpretation (Person-Mean Across All Variants)

Fig. 9. The indirect effect of PWS framing on message resistance and message effectiveness through equalizing interpretation (Study 4a).
Note. Although not depicted in Fig. 9 for simplicity, building-block interpretation (i.e., the person-centered score and person-mean) was also regressed onto the outcomes within the model. †$p < .10$, * $p < .05$, ** $p < .01$, *** $p < .001$.

Resistance did not differ between the standard PWS and the differentiation-short PWS ($b = -0.04, p = .101$, 95% CI [−0.09, 0.008]), or the differentiation-ladder PWS ($b = -0.01, p = .775$, 95% CI [−0.08, 0.06]). Anti-racist motivation did not differ between the standard PWS and the differentiation-short PWS ($b = 0.01$, $p = .762$, 95% CI [−0.03, 0.05]), or the Standard PWS and the differentiation-ladder ($b = -0.05, p = .082$, 95% CI [−0.10, 0.006]).

Mediation Analysis. Does reduced equalization account for modification effectiveness? Using a multi-level SEM path-model (Fig. 9) we tested whether reductions in equalization accounted for why the differentiation-long version of the PWS (relative to the standard PWS) received less resistance and was more effective in motivating anti-racist action. Given the within-person design of Study 4a, the multilevel mediation model had a 1–1-1 structure, such that we tested whether individual variation in equalizing interpretation for each of the four message variants (level 1) explained differences in individuals' resistance and motivation response to each message variant (level 1). To test this model, it was necessary to disentangle the effect of *within-person variation* in people's equalizing interpretation across the different variants on outcomes (our mediation pathway), from the between-person effect of people who are generally high in equalizing interpretation being more likely to resist all messages. To assess the within-person effect of people's equalizing interpretation, we centered participant's equalizing interpretation of each message variant around their person-level mean of all four messages, and treated the person-centered score (i.e., the within-person effect) as our mediator, while controlling for person-level mean equalizing interpretation (i.e., the between-person effect) on both outcomes (see Bauer, Preacher, & Gil, 2006; Zhang, Zyphur, & Preacher, 2009). We also controlled for the effect of building-block interpretation—both the within-person effect (i.e., person centered score) and between person effect (i.e., person mean)—on outcomes. We estimated indirect effects using 10,000 Monte Carlo simulations (Rockwood & Hayes, 2017). We note that we cannot infer causal order of all paths in this model: While message type must precede equalization, message resistance and anti-racist motivation (i.e., people's perceptions did not determine what message they were shown), an alternative model in which message resistance and anti-racist motivation precedes equalization could also be possible since these outcomes were all measured at once.

As predicted, we found significant within-person effects of having an equalizing interpretation on both outcomes, such that people tended to show significantly greater resistance to and were significantly less motivated to engage in anti-racist action by the message variants that elicited the greatest equalizing interpretation (resistance: $b = 0.32$, $p < .001$, 95% CI [0.24, 0.40]); anti-racist action: $b = -0.19$, $p < .001$, 95% CI [−0.25, −0.13).[16] This within-person effect of equalization partly mediated the effect of the differentiation-long PWS (versus standard PWS) on reduced message resistance (*Indirect Effect = 95%* MCCI [−0.10, −0.04]) and increased motivation to engage in anti-racist action (*Indirect Effect = 95%* MCCI [0.02, 0.06]; see Fig. 9 for all direct effects).

11.1.3. Discussion

We successfully reduced resistance to the PWS and increased the message's effectiveness in motivating anti-racist action when we made it more explicit that the different elements which contribute to structural racism (e.g., passive indifference versus racial violence) are distinct but causally related to each other. This modification did not reduce the effectiveness of the PWS message in communicating the structural

[16] We also tested the between and within person effects of equalizing interpretation for message resistance and anti-racist motivation in separate multilevel regressions controlling for the within and between effects of building-block interpretation and message type. We found significant within and between person effects of equalization for both message resistance (positive association) and anti-racist motivation (negative association); see Supplemental Table 29.

relation between indifference and active racism (i.e., building-block interpretation), nor did it undercut the perceived harm of passive indifference. There was nuance however: It was necessary to ensure White Americans that being indifferent does not have the same implications for one's moral character as blatant acts of racism—merely stating that indifference and active racism are distinct (as in the "short" version) was insufficient.

11.2. Study 4b

In Study 4b we sought to replicate the findings of Study 4a in the context of the "White Silence is White Violence" message.

11.2.1. Method

Procedure. We showed White Americans the standard "White Silence is White Violence" message, and three alternative versions that more clearly differentiated passive indifference from active racism. One message variant was "White Silence is a foundation for White Violence", which maintained the "is" statement, while eliminating the direct 'silence is violence' equation language by explicitly stating the building-block idea that silence is a foundation for violence. A second variant was "White Silence Contributes to White Violence". In this message we removed the "is" statement (which might be threatening), but still explicitly stated that silence plays a causal role in contributing to violence by using agentive transitive language (Fausey & Boroditsky, 2010). Finally, a third message variant was "Ending White Silence Can Help End White Violence". In this variant we removed the agentive transitive language by switching the emphasis from the negative "Silence" verb to the positive "Ending Silence" verb. Thus, this variant may be met with less resistance because it does not directly imply that people who are silent contribute to violence (Fausey & Boroditsky, 2010). A possible limitation of this variant however is that it speaks less directly to the harm of indifference relative to the other variants.

Participants. We recruited 451 White Americans from Prolific completed the study between February 17th, 2021 and February 24th, 2021. We determined sample size and concluded data collection prior to analyses. After pre-registered exclusions conducted prior to analyses (aspredicted.org/blind.php?x=/DQL_BOP), our final sample consisted of 370 White Americans ($M_{age} = 36.80$; $SD_{age} = 13.43$; 208 Female, 162 Male). This sample size yielded 1480 observations (because of repeated-measures design) ensuring at least 5–10 observations per parameter for our planned SEM analyses (in this case at least an $n = 120$; see Kline, 2011).

Measures. We used the same items as Study 4a (but adapted to the "White Silence" message) to measure *equivalizing interpretation* ($\alpha_{standard} = 0.75$; $\alpha_{foundation} = 0.83$; $\alpha_{contributes} = 0.82$; $\alpha_{end\_silence} = 0.81$), *building-block interpretation* ($\alpha_{standard} = 0.86$; $\alpha_{foundation} = 0.80$; $\alpha_{foundation} = 0.84$; $\alpha_{end\_silence} = 0.85$), harm of White Silence (one item), message resistance (one item), and anti-racist motivation (two items; $r_{standard} = 0.96$; $r_{\_foundation} = 0.97$; $r_{\_contributes} = 0.97$; $r_{end\_silence} = 0.97$).[17]

Antecedents of Equalizing Interpretation. We also assessed potential antecedents of having an equalizing interpretation: resistance to environmental heuristics, collective narcissism, ethnic identification, political conservatism, and racial anxiety. Replicating Study 1 and Study 4a, collective narcissism was positively associated with greater equalization across the different message variants. However, unlike Study 1 and Study 4a, racial anxiety and political conservatism was not associated with equalization. The relation between having an equalizing interpretation and ethnic identification was non-significant as in previous studies (See Supplemental Tables 35–37 for detailed results).

11.2.2. Results

Descriptive statistics and Pearson-Correlations are summarized in Supplemental Tables 33–34. Consistent with Study 4a, we used a multilevel SEM path model to test the effect of each "White Silence" message variant (relative to the standard message) on all outcome measures (see Study 4b supplemental information for ICCs of nested variables). Mean ratings for all outcomes as a function of variant type are summarized in Table 12.[18]

11.2.2.1. Is it possible to reduce having an equalizing interpretation without compromising having a building-block interpretation?

Equalizing Interpretation. Compared to the standard message, the "White Silence is a foundation for White Violence" message ($b = -1.02$, $p < .001$, 95% CI [−1.12 -0.92]), the "White Silence contributes to White Violence" message ($b = -0.97$, $p < .001$, 95% CI [−1.07, −0.86]), and the "Ending White Silence can help end White Violence" message ($b = -1.19$, $p < .001$, 95% CI [−1.31, −1.08]) elicited significantly less equalizing interpretation.

Building-Block Interpretation. The "White Silence is a foundation for White Violence" message ($b = 0.29$, $p < .001$, 95% CI [0.19, 0.38]), and the "White Silence contributes to White Violence" message ($b = 0.16$, $p = .002$, 95% CI [0.06, 0.26]) elicited significantly more building-block interpretation than the standard message. There were no differences in building-block interpretation between the standard message and the "Ending White Silence can help end White Violence" message ($b = 0.02$, $p = .799$, 95% CI [−0.10, 0.13]).

Harm of indifference. The "White Silence is a foundation for White Violence" message ($b = 0.07$, $p = .118$, 95% CI [−0.02, 0.16]), the "White Silence contributes to White Violence" message ($b = 0.02$, $p = .620$, 95% CI [−0.07, 0.12]), and the "Ending White silence can help end White Violence" message ($b = -0.10$, $p = .068$, 95% CI [−0.21, 0.007]) were all just as effective as the standard "White Silence is White Violence" message in communicating the harm of silence.

Key Outcomes. Compared to the standard White Silence message, the "White Silence is a foundation for White Violence" message ($b = -0.46$, $p < .001$, 95% CI [−0.55, −0.38]), the " White Silence contributes to White Violence" message ($b = -0.42$, $p < .001$, 95% CI [−0.50, −0.32]), and the "Ending White Silence can help end White Violence" message ($b = -0.48$, $p < .001$, 95% CI [−0.58, −0.38]) were all significantly less likely to evoke message resistance.

Compared to the standard White Silence message, the "White Silence is a foundation for White Violence" message ($b = 0.20$, $p < .001$, 95% CI [0.14, 0.26]), the " White Silence contributes to White Violence" message ($b = 0.20$, $p < .001$, 95% CI [0.13, 0.27]), and the "Ending White Silence can help end White Violence" message ($b = 0.20$, $p < .001$, 95% CI [0.12, 0.28]) were all rated by White Americans as being more effective in motivating them to be anti-racist.

Mediation Analysis. Does reduced equalization account for the effectiveness of the modified messages? Using a multi-level SEM path model following the same approach used in Study 4a (see Fig. 10), we tested whether reductions in equalizing interpretation accounted for why the modified "White Silence" messages received less resistance and motivated greater anti-racist action intentions compared to the standard message (controlling for the effect of participants' building-block interpretation) on each outcome (See Study 4b Supplemental Information for model details). For the same reasons described in Study 4a, we cannot rule out the possibility that message resistance and anti-racist

[17] We initially pre-registered that we would use a difference score between equalization and building-block interpretation (i.e., the relative tendency for people to hold an equalization versus building-block interpretation) in our main text analyses. However, we now report analyses using the difference score in Supplemental Analyses (see Supplemental Tables 38, 40, 41, and Supplemental Fig. 11) and treat equalization and building-block interpretation as separate variables in our main text analysis.

[18] We note that results remain consistent when analyzing each outcome separately with independent multilevel regression models or when using a repeated measures ANOVA to test differences between variants.

Table 12
Descriptive statistics for all measured variables by message type (Study 4b).

	"Silence Is Violence"		"A foundation for Violence"		"Contributes to Violence"		"End Silence End Violence"	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
1. Equalizing Interpretation	5.38^{a}	1.41	3.58^{b}	1.64	3.67^{b}	1.57	3.27^{c}	1.63
2.Building-Block Interpretation	5.47^{a}	1.41	5.83^{b}	1.12	5.67^{c}	1.19	5.48^{a}	1.32
3. Harm of Indifference	4.32^{a}	1.53	4.43^{a}	1.4	4.36^{a}	1.41	4.18^{a}	1.47
4. Message Resistance	3.65^{a}	1.74	2.86^{b}	1.66	2.94^{b}	1.62	2.84^{b}	1.66
5. Anti-Racist Motivation	3.50^{a}	1.74	3.84^{b}	1.65	3.84^{b}	1.65	3.83^{b}	1.65

Note. Within a row, means with different subscripts are significantly different ($p < .05$).

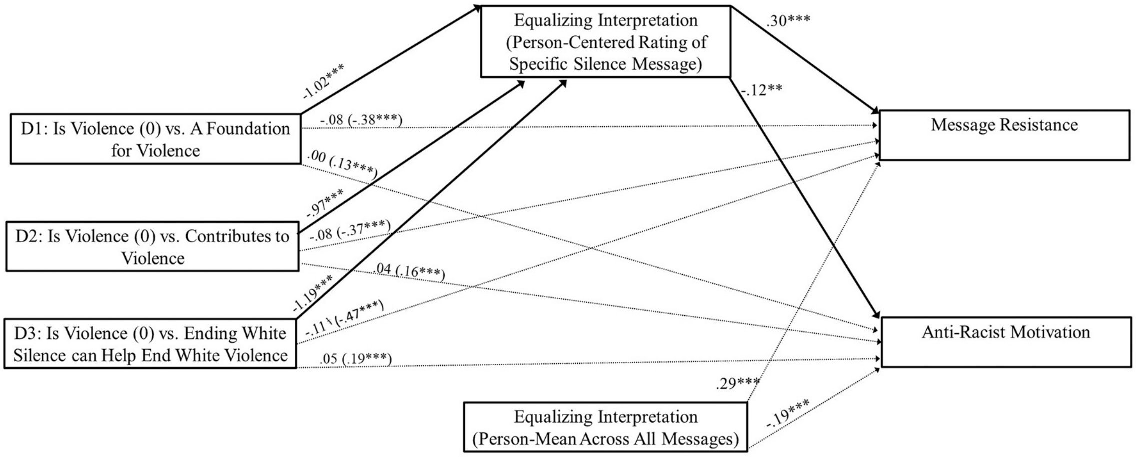

Equalizing Interpretation
(Person-Centered Rating of
Specific Silence Message)
D1: Is Violence (0) vs. A Foundation for Violence
D2: Is Violence (0) vs. Contributes to Violence
D3: Is Violence (0) vs. Ending White Silence can Help End White Violence
Message Resistance
Anti-Racist Motivation
Equalizing Interpretation
(Person-Mean Across All Messages)
-1.02***
-.97***
-1.19***
.30***
-.12**
-.08 (-.38***)
.00 (.13***)
-.08 (-.37***)
.04 (.16***)
-.11\ (-.47***)
.05 (.19***)
.29***
-.19***

Fig. 10. The indirect effect of "white silence" message framing on message resistance and message effectiveness through equalizing interpretation (Study 4b). *Note.* Although not depicted in Fig. 10 for simplicity, building-block interpretation (i.e., the person-centered score and person-mean) were also regressed onto the outcomes within the model. Total effects reported in parentheses. \ $p < .10$, * $p < .05$, ** $p < .01$, *** $p < .001$.

motivation precedes equalization because these outcomes were all measured at once.

As predicted, we found significant within-person effects of equalizing interpretation on both outcomes: People were significantly more likely to resist ($b = 0.30$, $p < .001$, 95% CI [0.21, 0.39]) and were significantly less motivated to engage in anti-racist action by ($b = -0.12$, $p = .006$, 95% CI [−0.21, −0.04) the message variants they interpreted to be the highest in equalizing interpretation.[19] Reduced equalizing interpretation accounted in part for why the "White Silence is a foundation for White Violence" message (*Indirect Effect* = *95%* MCCI [−0.41, −0.21]), the "White Silence contributes to White Violence" message (*Indirect Effect* = *95%* MCCI [−0.39, −0.20]), and the "Ending White Silence can help end White Violence" message (*Indirect Effect* = *95%* MCCI [−0.47, −0.25]) all received less resistance than the standard message. Similarly, reduced equalizing interpretation accounted in part for why the "White Silence is a foundation for White Violence" message (*Indirect Effect* = *95%* MCCI [0.04, 0.22]), the "White Silence contributes to White Violence" message (*Indirect Effect* = *95%* MCCI [0.04, 0.21]), and the "Ending White Silence can help end White Violence" message (*Indirect Effect* = [0.04, 0.25]) were more effective in motivating anti-racist action intentions than the standard message.

11.2.3. Discussion

We replicated Study 4a in the context of the "White Silence is White Violence" message. We successfully reduced message resistance and increased the message's effectiveness in motivating anti-racist action by reducing White American's equalizing interpretation via small wording changes (i.e., avoiding "silence *is* violence" in favor of saying that silence is "a foundation for violence" or "contributes to violence"). These modifications did not reduce the message's effectiveness in communicating a building-block interpretation; indeed, it amplified it. Nor did these modifications undercut the perceived harm of silence.

12. General discussion

Four studies with large samples (one nationally representative) reveal that different interpretations of anti-racist messages about structural racism (i.e., those that emphasize the need to actively dismantle structural racial inequities through anti-racist action; Kendi, 2019; Ansley, 1997) contribute to their divisiveness. White Americans with an "equalizing" interpretation—believing that anti-racist messages equate indifference with violence—were more opposed to these messages compared with those with a "building-block" interpretation—believing that indifference helps facilitate and maintain downstream violence and inequity (Study 1). Having an equalizing interpretation was most common in White Americans highest in collective narcissism and racial anxiety (although only collective narcissism was a robust predictor across all studies). Experimentally inducing White Americans to hold an equalizing interpretation of anti-racist messages also increased their level of identity-based threat and their resistance to anti-racist policies (Study 2). Differences in interpretation also shaped the effect of seeing anti-racist messages, with people high in equalizing interpretation and low in building-block interpretation responding upon message exposure with increased identity threat and denial of ongoing anti-black discrimination (Study 3). Importantly, it was possible to reduce this divisiveness and backlash without decreasing the effectiveness of messaging about structural racism—we did so by

[19] We also tested the between and within person effects of equalization for message resistance and anti-racist motivation in separate multi-level regressions controlling for the within and between effects of building-block interpretation and message type. We found significant within and between person effects of equalization for both message resistance (positive association) and anti-racist motivation (negative association); see Supplemental Table 39.

nudging people towards a building-block (vs. equalizing) interpretation (Study 4).

12.1. Implications

Mounting research highlights both the benefits and barriers of promoting White Americans to confront structural racism (Adams et al., 2008; Bonam et al., 2019; Rucker et al., 2019; Rucker & Richeson, 2021; Salter et al., 2018), and racial inequities (Kraus, Rucker, & Richeson, 2017; Kraus, Torrez, & Hollie, 2022; Lowery et al., 2007; Onyeador et al., 2021; Phillips & Lowery, 2015). This past work reveals how the idea of structural racism and White privilege is generally threatening to White Americans, and offers psychological strategies for how White Americans might cope with such threat. For example, White Americans who affirm other parts of their White identity (Gunn & Wilson, 2011; Lowery et al., 2007; Unzueta & Lowery, 2008) or engage in emotion regulation strategies (Ford et al., 2022) might be better able to cope with the threat of acknowledging how they are privileged by structural inequities. We do not dispute that confronting structural racism is challenging to some extent for all White Americans. However, our theoretical approach is distinct from past work in that we consider psychological factors that might make these messages appear more threatening to some White Americans than others. Specifically, we position equalizing versus building-block interpretations of structural racism as a novel factor involved in why some White Americans choose to deny (versus dismantle) structural inequities (Knowles et al., 2014). Our work also suggests that equalizing versus building-block interpretations may be one factor contributing to why critical race theory (which emphasizes the need for White Americans to actively challenge systems of racial oppression; Kendi, 2019; Ansley, 1997) is so contentious in the United States (Ray & Gibbons, 2021). Yet as we discuss below, White Americans' fears of status loss and prejudice likely also contributed to backlash against CRT.

Our introduction of equalizing vs. building-block interpretations as a potential source of threat and backlash suggests a new strategy for modifying anti-racist messages about structural racism to evoke less threat and resistance—framing these messages to elicit a building-block interpretation, not an equalizing interpretation. Our approach might be particularly useful when the objective is to garner White Americans' support for anti-racism when they are exposed to these messages on social media (Capatides, 2020), on the news (Fox News, Jan 18th, 2018), or at work (Pothast, 2021). In such contexts there might not be any anti-racism facilitator to guide White people in processing their response via other strategies such as identity affirmation or emotion regulation, and thus the message itself might be the only vehicle for intervention available.

We acknowledge that the approach of modifying the anti-racist message puts the onus on the message creator (rather than the White recipient of the message) to put in the work to minimize potential threat. And indeed, it is critical for White individuals to play their part in embracing the discomfort of confronting structural racism (Kleine, 2018) by using strategies like self-affirmation (Unzueta & Lowery, 2008) or emotion regulation (Ford et al., 2022). It is also important to consider that the optimal framing of an anti-racist message depends on the context it is being shown and the objectives for showing the message. For example, during Black Lives Matter protests the primary objective of the "White Silence" message might be to draw people's immediate attention to the issue of structural racism or to express the psychological pain that Black communities experience when White communities remain silent about structural racism (Capatides, 2020)—in such a context the original "White silence is violence" message might, on account of its pithiness, be optimal despite its potential to elicit an equalizing interpretation. Thus, a challenge of message creators may be to balance the goal of catching people's attention and powerfully conveying the importance of challenging structural racism while also managing the feelings the message might elicit among both potential supporters and detractors of the message (also see Kendi, 2019; chapter 16).

Our research builds on past work that considers the psychological processes involved when White people receive formal education about structural racism (Tatum, 1992). While some messages about structural racism like the PWS were initially designed for use in structured anti-racism training workshops, it is a reality that White Americans often encounter these messages on social media, the news, or their company work channel when there is no facilitator present to guide interpretation. Our research provides insight as to how White individuals naturally respond to real-world anti-racist messages without any external guidance or intervention. Still it is important to consider how the present research findings relate to approaches that trained facilitators might employ when guiding critical conversations about structural racism. For example, critical race scholars emphasize the importance of shifting the White person's focus from the question of whether or not they are racist in character (something stable and morally threatening) to how to they might avoid behaviors that reinforce racist systems (something malleable and under the person's control; Kendi, 2019; Tatum, 1992). Such strategies might reduce the extent to which White people have an equalizing interpretation of messages like the PWS by shifting White people's focus away from the question of whether being silent makes them a blatant racist. An exciting direction for future research will be to examine how equalizing versus building-block interpretations of structural racism change among White individuals as they undergo formal anti-racism trainings.

Our approach of focusing on real-world messages about structural racism also fills a void in prejudice reduction research that typically tests the effects of new prejudice-reduction materials based on abstract psychological theories, rather than the materials used in the real-world (Paluck, Porat, Clark, & Green, 2021). Moreover, our work is distinct from the majority of prejudice reduction research, which typically focuses on reducing individual racial biases (i.e., training people to not be racist) rather than on motivating people to challenge structural inequities (i.e., being anti-racist; Kendi, 2019; Paluck et al., 2021; but see Adams et al., 2008; Bonam et al., 2019; Rucker et al., 2019; Rucker & Richeson, 2021).

Still, useful connections can be drawn between our research, and research focusing on the role of moral threat in White Americans' willingness to acknowledge individual (implicit) biases (Daumeyer, Onyeador, Brown, & Richeson, 2019; Vitriol & Moskowitz, 2021). Research suggests that White individuals might resist acknowledging that they hold implicit racial biases if they feel they will be morally condemned as racist for holding those implicit biases in the same way that they would be for expressing explicit prejudice (Vitriol & Moskowitz, 2021). Based on our research findings, it is likely that equalizing interpretations might contribute to these feelings of threat. Yet, other research suggests that some White individuals may view implicit versus explicit racial biases as less morally blameworthy and thus less in need of intervention (Daumeyer et al., 2019). White individuals low in equalizing interpretation might feel little urgency to confront implicit racial biases if they view such acts less morally wrong than explicit biases. As we have shown in our work however, being low in equalizing interpretation does not mean that people view the different behaviors that contribute to racist systems as more or less harmful or morally acceptable – having a building-block interpretation still acknowledges that more mild acts like silence or implicit biases play a fundamental role in maintaining racist systems and are thus morally unacceptable. For example, our results of Study 4 suggested that by decreasing White Americans' level of equalizing interpretation by modifying anti-racist messages, we did not reduce the message's effect in conveying the harm of White silence and the necessity for intervention. While harm perceptions are closely linked to perceptions of immorality (Schein & Gray, 2018), future research is needed to ensure that decreasing White people's equalizing interpretation does not lead them to view White silence as more morally acceptable.

12.2. Limitations and future directions

We presented messages about structural racism using a "light-touch" approach where participants only briefly reflected on the meaning of these messages. This approximates how people see these messages quickly on social media or organizational message feeds. However, "light-touch" designs often have small effects (Paluck et al., 2021). Although we found that being high in equalizing interpretation and low in building-block interpretation was robustly associated with White identity threat and backlash across all studies, we found that experimentally being shown standard versions of anti-racist messages about structural racism (vs. no exposure; Study 3) generally had little effect on White Americans. Only White Americans high in equalizing interpretation and low in building-block interpretation tended to show small increases in White identity threat and subsequent denial of racism. Future work should explore whether larger effects might emerge when White Americans can also engage in discussion and potential debate about these messages with other White Americans (as might occur when individuals encounter these messages on social media or in other group settings). Future work should also consider whether anti-racist messages about structural racism can have larger positive effects in structured anti-racism seminars (Adams et al., 2008; Tatum, 1992) where trained facilitators are present to walk individuals through the meaning of structural racism and compassionately rebut concerns about equalization.

The present work is limited to online convenience samples and relied on one-time self-report measures. While we recruited a representative sample in Study 1, future research should take a field approach, examining how different permutations of messages about structural racism impact how people in classrooms or organizations confront structural racism and actually behave. Moreover, it will be important for future research to include a broader range of outcomes – for instance, future work is needed to test the effect of different message variations on actual anti-racist behavior (e.g., writing a letter to a state governor). As well, further work should test whether the novel framings developed in Study 4 might also function to decrease accountability for less overt (silence) versus more overt (racist jokes) behaviors that contribute to racist systems (Daumeyer et al., 2019). However, given that several of the intervention message framings tested in Study 4 did not reduce White American's perception that silence/inaction is harmful, we suspect that the re-framed messages tested in this work would not reduce accountability.

Study 1 relied on correlational data, and therefore left open questions regarding the causal order of whether White identity threat and backlash result from or precede equalization. However, in Study 2 we partly addressed these concerns by examining whether experimentally inducing equalizing interpretations increased threat and defensive responding. Further, in Study 3 we found that showing people standard messages (versus no message) affected their downstream attitudes, depending on their tendency to hold an equalizing and building-block interpretation.

Finally, it remains an open question whether White people's differing interpretations of anti-racist messages results from basic processing differences in how they interpret these messages, and/or because of motivated processes that lead some White people to view these messages in a negative light to legitimize the maintenance of racial inequities that privilege White Americans (Chow, Lowery, & Hogan, 2013; Knowles, Lowery, & Schaumberg, 2009; Kteily & McClanahan, 2020; Lowery, Unzueta, Knowles, & Goff, 2006). Such motivated processes could be due to strategic choices to view the message in a threatening way: That is, rather than honestly interpreting a message as equalizing and being threatened by it, certain individuals might cynically frame a message as equalizing even when they know it isn't, simply because they know an equalizing message is likely to stoke fear in others and help to justify their resistance to change. It is also possible that these motivated processes to maintain hierarchy could be acting outside of conscious awareness. Although we found that White collective narcissism was robustly positively associated with greater equalizing interpretation, the correlational nature of this link makes it unclear whether White people who view Whites as important and deserving of status (Marchlewska et al., 2020) encode anti-racist messages as more threatening at a basic cognitive level or are motivated to view these messages negatively to protect their status. Future experimental research could test whether equalizing interpretations result from a motivated processes by examining whether White people report a greater equalizing interpretation of the same anti-racist message when their status is threatened (versus made secure; Craig & Richeson, 2014).

13. Conclusion

Anti-racist messages inspired by Critical Race Theory about structural racism evoke heated debate, but they are important for explaining how racist systems can be propagated even if the majority of people strongly value racial equity. Our work suggests that one way to bring Americans together around these messages is to emphasize that even seemingly insignificant behaviors can facilitate and build up to highly problematic ones—while also emphasizing that these behaviors are distinct.

Authors' contributions

Kachanoff, Kteily and Gray developed the research ideas together and formulated the research design. Kachanoff took the lead in conducting the studies and analyzed the data. Kachanoff wrote the initial draft of the paper and performed the revisions. Kteily and Gray helped revise the initial paper draft and gave feedback throughout the writing process and revision process.

All data, analysis scripts, and materials shown to participants are available on the OSF: https://osf.io/jr2t4/?view_only=8638649403294319a9d20e7097ddf123.

Funding Sources

Charles Koch Foundation - Center for the Science of Moral Understanding.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jesp.2022.104348.

References

Adams, G., Edkins, V., Lacka, D., Pickett, K., & Cheryan, S. (2008). Teaching about racism: Pernicious implications of the standard portrayal. *Basic and Applied Social Psychology, 30*, 349–361. https://doi.org/10.1080/01973530802502309

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

Ansley, F. L. (1997). White supremacy (and what we should do about it). In R. Delgado, & J. Stefancic (Eds.), *Critical white studies: Looking behind the Mirror* (pp. 592–655). Philadelphia, PA: Temple University Press.

Asare, J. G. (2020). *Merriam-Webster is changing the definition of racism to reflect systemic oppression*. Forbes. https://www.forbes.com/sites/janicegassam/2020/06/11/merriam-webster-is-changing-the-definition-of-racism-to-reflect-systemic-oppression/?sh=479200b0400f.

Bauer, D. J., Preacher, K. J., & Gil, K. M. (2006). Conceptualizing and testing random indirect effects and moderated mediation in multilevel models: New procedures and recommendations. *Psychological Methods, 11*, 142–163. https://doi.org/10.1037/1082-989X.11.2.142

Blake, K. R., & Gangestad, S. (2020). On attenuated interactions, measurement error, and statistical power: Guidelines for social and personality psychologists. *Personality and Social Psychology Bulletin, 46*, 1702–1711. https://doi.org/10.1177/0146167220913363

Bonam, C. M., Nair Das, V., Coleman, B. R., & Salter, P. (2019). Ignoring history, denying racism: Mounting evidence for the Marley hypothesis and epistemologies of ignorance. *Social Psychological and Personality Science, 10*, 257–265. https://doi.org/10.1177/1948550617751583

Byrne, B. M. (1994a). *Structural equation modeling with EQS and EQS/Windows: Basic concepts, applications, and programming*. CA: Sage: Newbury Park.

Byrne, B. M. (1994b). Testing for the Factorial Validity, Replication, and Invariance of a Measuring Instrument: A Paradigmatic Application Based on the Maslach Burnout Inventory. *Multivariate Behavioral Research, 29*(3), 289–311. https://doi.org/10.1207/s15327906mbr2903_5

Capatides, C. (2020). White silence on social media: Why not saying anything is actually saying a lot. *CBS News*. https://www.cbsnews.com/news/white-silence-on-social-media-why-not-saying-anything-is-actually-saying-a-lot/.

Carpenter, S. (2018). Ten steps in scale development and reporting: A guide for researchers. *Communication Methods and Measures, 12*, 25–44. https://doi.org/10.1080/19312458.2017.1396583

Chow, R. M., Lowery, B. S., & Hogan, C. M. (2013). Appeasement: Whites' strategic support for affirmative action. *Personality and Social Psychology Bulletin, 39*, 332–345. https://doi.org/10.1177/0146167212475224

Clifford, S., Sheagley, G., & Piston, S. (2021). Increasing precision without altering treatment effects: Repeated measures designs in survey experiments. *American Political Science Review, 115*, 1048–1065. https://doi.org/10.1017/S0003055421000241

Costafreda, S. G. (2009). Pooling FMRI data: meta-analysis, mega-analysis and multi-center studies. *Frontiers in Neuroinformatics, 3*, 33. https://doi.org/10.3389/neuro.11.033.2009

Craig, M. A., & Richeson, J. A. (2014). More diverse yet less tolerant? How the increasingly diverse racial landscape affects white Americans' racial attitudes. *Personality and Social Psychology Bulletin, 40*, 750–761. https://doi.org/10.1177/0146167214524993

Curran, P. J., & Hussong, A. M. (2009). Integrative data analysis: The simultaneous analysis of multiple data sets. *Psychological Methods, 14*, 81. https://doi.org/10.1037/a0015914

Daumeyer, N. M., Onyeador, I. N., Brown, X., & Richeson, J. A. (2019). Consequences of attributing discrimination to implicit vs. explicit bias. *Journal of Experimental Social Psychology, 84*, Article 103812. https://doi.org/10.1016/j.jesp.2019.04.010

DiAngelo, R. (2018). *White fragility: Why it's so hard for white people to talk about racism.* Boston, Massachusetts: Beacon Press.

Dovidio, J. F., & Gaertner, S. L. (Eds.). (1986). *Prejudice, discrimination, and racism* (pp. 61–89). Orlando, FL: Academic Press.

Fausey, C. M., & Boroditsky, L. (2010). Subtle linguistic cues influence perceived blame and financial liability. *Psychonomic Bulletin & Review, 17*, 644–650. https://doi.org/10.3758/PBR.17.5.644

Fiedler, K., Harris, C., & Schott, M. (2018). Unwarranted inferences from statistical mediation tests–an analysis of articles published in 2015. *Journal of Experimental Social Psychology, 75*, 95–102. https://doi.org/10.1016/j.jesp.2017.11.008

Ford, B. Q., Green, D. J., & Gross, J. J. (2022). White fragility: An emotion regulation perspective. *The American Psychologist*. https://doi.org/10.1037/amp0000968

Fox News. (2018). 'Pyramid of white supremacy' used in university curriculum. *Fox News*. https://video.foxnews.com/v/5713923343001#sp=show-clips.

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives, 19*, 25–42. https://doi.org/10.1257/089533005775196732

Gillborn, D. (2006). Rethinking white supremacy: Who counts in 'white world'. *Ethnicities, 6*, 318–340. https://doi.org/10.1177/1468796806068323

Ginersorolla, R. (2018). Powering your interaction. *Approaching Significance*. https://approachingblog.wordpress.com/2018/01/24/powering-your-interaction-2/.

Golec de Zavala, A, Cichocka, A., Eidelson, R., & Jayawickreme, N. (2009). Collective narcissism and its social consequences. *Journal of Personality and Social Psychology, 97*, 1074–1096. https://doi.org/10.1037/a0016904

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review, 102*, 4–27.

Gunn, G. R., & Wilson, A. E. (2011). Acknowledging the skeletons in our closet: The effect of group affirmation on collective guilt, collective shame, and reparatory attitudes. *Personality and Social Psychology Bulletin, 37*, 1474–1487. https://doi.org/10.1177/0146167211413607

Hansson, D. H. (2021, April 28). Let it all out. *Personal Blog Post*. https://world.hey.com/dhh/let-it-all-out-78485e8e.

Hayes, A. F. (2006). A primer on multilevel modeling. *Human Communication Research, 32*, 385–410. https://doi.org/10.1111/j.1468-2958.2006.00281.x

Ho, A. K., Sidanius, J., Kteily, N., Sheehy-Skeffington, J., Pratto, F., Henkel, K. E., … Stewart, A. L. (2015). The nature of social dominance orientation: Theorizing and measuring preferences for intergroup inequality using the new SDO_7 scale. *Journal of Personality and Social Psychology, 109*, 1003–1028. https://doi.org/10.1037/pspi0000033

Jost, J. T., Nosek, B. A., & Gosling, S. D. (2008). Ideology: Its resurgence in social, personality, and political psychology. *Perspectives on Psychological Science, 3*, 126–136. https://doi.org/10.1111/j.1745-6916.2008.00070.x

Kachanoff, F. J., Kteily, N. S., Khullar, T. H., Park, H. J., & Taylor, D. M. (2020). Determining our destiny: Do restrictions to collective autonomy fuel collective action? *Journal of Personality and Social Psychology, 119*, 600–632. https://doi.org/10.1037/pspi0000217

Kachanoff, F. J., Taylor, D. M., Caouette, J., Khullar, T. H., & Wohl, M. J. A. (2019). The chains on all my people are the chains on me: Restrictions to collective autonomy undermine the personal autonomy and psychological well-being of group members. *Journal of Personality and Social Psychology, 116*, 141–165. https://doi.org/10.1037/pspp0000177

Kendi, I. X. (2019). *How to be an antiracist*. New York, New York: One World.

Kleine, E (2018). White threat in a browning America. *Vox*. https://www.vox.com/policy-and-politics/2018/7/30/17505406/trump-obama-race-politics-immigration.

Kline, R. (2011). Convergence of structural equation modeling and multilevel modeling. *The SAGE handbook of innovation in social research methods* (pp. 562–589). SAGE Publications Ltd. https://doi.org/10.4135/9781446268261

Knowles, E. D., Lowery, B. S., Chow, R. M., & Unzueta, M. M. (2014). Deny, distance, or dismantle? How white Americans manage a privileged identity. *Perspectives on Psychological Science, 9*, 594–609. https://doi.org/10.1177/1745691614554658

Knowles, E. D., Lowery, B. S., & Schaumberg, R. L. (2009). Anti-egalitarians for Obama? Group-dominance motivation and the Obama vote. *Journal of Experimental Social Psychology, 45*, 965–969. https://doi.org/10.1016/j.jesp.2009.05.005

Kraus, M. W., Rucker, J. M., & Richeson, J. A. (2017). Americans misperceive racial economic equality. *Proceedings of the National Academy of Sciences, 114*(39), 10324–10331. https://doi.org/10.1073/pnas.1707719114

Kraus, M. W., Torrez, B., & Hollie, L. (2022). How narratives of racial progress create barriers to diversity, equity, and inclusion in organizations. *Current Opinion in Psychology, 43*, 108–113. https://doi.org/10.1016/j.copsyc.2021.06.022

Kteily, N. S., & McClanahan, K. J. (2020). Incorporating insights about intergroup power and dominance to help increase harmony and equality between groups in conflict. *Current Opinion in Psychology, 33*, 80–85. https://doi.org/10.1016/j.copsyc.2019.06.030

Leach, C. W., Van Zomeren, M., Zebel, S., Vliek, M. L., Pennekamp, S. F., Doosje, B., … Spears, R. (2008). Group-level self-definition and self-investment: A hierarchical (multicomponent) model of in-group identification. *Journal of Personality and Social Psychology, 95*, 144–165. https://doi.org/10.1037/0022-3514.95.1.144

Ledgerwood, A., Hudson, S. T. J., Lewis, N. A., Jr., Maddox, K. B., Pickett, C. L., Remedios, J. D., … Wilkins, C. L. (2022). The pandemic as a portal: Reimaging psychological science as truly open and inclusive. *Perspectives on Psychological Science*. https://doi.org/10.1177/17456916211036654

Lowery, B. S., Knowles, E. D., & Unzueta, M. M. (2007). Framing inequity safely: Whites' motivated perceptions of racial privilege. *Personality and Social Psychology Bulletin, 33*, 1237–1250. https://doi.org/10.1177/0146167207303016

Lowery, B. S., Unzueta, M. M., Knowles, E. D., & Goff, P. A. (2006). Concern for the in-group and opposition to affirmative action. *Journal of Personality and Social Psychology, 90*, 961–974. https://doi.org/10.1037/0022-3514.90.6.961

Marchlewska, M., Cichocka, A., Jaworska, M., Golec de Zavala, A., & Bilewicz, M. (2020). Superficial ingroup love? Collective narcissism predicts ingroup image defense, outgroup prejudice, and lower ingroup loyalty. *British Journal of Social Psychology, 59*, 857–875. https://doi.org/10.1111/bjso.12367

McGivney, A. (2021). The battle for Mount Rushmore: "It should be turned into something like the holocaust museum". *The Guardian*. https://www.theguardian.com/environment/2021/jul/03/mount-rushmore-south-dakota-indigenous-americans.

Onyeador, I. N., Daumeyer, N. M., Rucker, J. M., Duker, A., Kraus, M. W., & Richeson, J. A. (2021). Disrupting beliefs in racial progress: Reminders of persistent racism alter perceptions of past, but not current, racial economic equality. *Personality and Social Psychology Bulletin, 47*, 753–765. https://doi.org/10.1177/0146167220942625

Paluck, E. L., Porat, R., Clark, C. S., & Green, D. P. (2021). Prejudice reduction: Progress and challenges. *Annual Review of Psychology, 72*, 533–560. https://doi.org/10.1146/annurev-psych- 071620-030619

Peetz, J., Gunn, G. R., & Wilson, A. E. (2010). Crimes of the past: Defensive temporal distancing in the face of past in-group wrongdoing. *Personality and Social Psychology Bulletin, 36*, 598–611. https://doi.org/10.1177/0146167210364850

Phillips, L. T., & Lowery, B. S. (2015). The hard-knock life? Whites claim hardships in response to racial inequity. *Journal of Experimental Social Psychology, 61*, 12–18. https://doi.org/10.1016/j.jesp.2015.06.008

Pothast, E. (2021). The "pyramid of hate" that brought down basecamp. *Marker*. https://marker.medium.com/the-pyramid-of-hate-that-brought-down-basecamp-838b63ca27e.

Ray, R., & Gibbons, A. (2021). Why are states banning critical race theory?. https://www.brookings.edu/blog/fixgov/2021/07/02/why-are-states-banning-critical-race-theory/.

Rockwood, N. J., & Hayes, A. F. (2017, May). MLmed: An SPSS macro for multilevel mediation and conditional process analysis. In *Poster presented at the annual meeting of the Association of Psychological Science (APS), Boston, MA*.

Rucker, J., Duker, A., & Richeson, J. (2019). *Structurally unjust: How lay beliefs about racism relate to perceptions of and responses to racial inequality in criminal justice*. https://doi.org/10.31234/osf.io/sjkeq

Rucker, J. M., & Richeson, J. A. (2021). Toward an understanding of structural racism: Implications for criminal justice. *Science, 374*, 286–290. https://doi.org/10.1126/science.abj7779

Salter, P. S., Adams, G., & Perez, M. J. (2018). Racism in the structure of everyday worlds: A cultural-psychological perspective. *Current Directions in Psychological Science, 27*, 150–155. https://doi.org/10.1177/0963721417724239

Schein, C., & Gray, K. (2018). The theory of dyadic morality: Reinventing moral judgment by redefining harm. *Personality and Social Psychology Review, 22*, 32–70. https://doi.org/10.1177/1088868317698288

Schmidt, S. L. (2005). More than men in white sheets: Seven concepts critical to the teaching of racism as systemic inequality. *Equity & Excellence in Education, 38*, 110–122. https://doi.org/10.1080/10665680590935070

Shnabel, N., & Nadler, A. (2015). The role of agency and morality in reconciliation processes: The perspective of the needs-based model. *Current Directions in Psychological Science, 24*, 477–483. https://doi.org/10.1177/0963721415601625

Simonsohn, U. (2014). No-way interactions. In *Data Colada*. http://datacolada.org/17.

sosspeace.org. (2019). Why our action matters: The Pyramid of White Supremacy. Retrieved August 6, 2021, from https://sosspeace.org/wp-content/uploads/2019/05/Appendix-1-Pyramid-of-White-Supremacy.pdf.

Sue, D. W. (Ed.). (2010). *Microaggressions and marginality: Manifestation, dynamics, and impact*. New York, NY: John Wiley & Sons.

Sullivan, D., Landau, M. J., Branscombe, N. R., & Rothschild, Z. K. (2012). Competitive victimhood as a response to accusations of ingroup harm doing. *Journal of Personality and Social Psychology, 102*, 778–795. https://doi.org/10.1037/a0026573

Takahashi, K., & Jefferson, H. (2021, February 4). *When the powerful feel voiceless: White identity and feelings of racial voicelessness*. https://doi.org/10.31234/osf.io/ry97q

Tatum, B. (1992). Talking about race, learning about racism: The application of racial identity development theory in the classroom. *Harvard Educational Review, 62*, 1–25. https://doi.org/10.17763/haer.62.1.146k5v980r703023

Thomson, K. S., & Oppenheimer, D. M. (2016). Investigating an alternate form of the cognitive reflection test. *Judgment and Decision making, 11*, 99–113.

Toplak, M. E., West, R. F., & Stanovich, K. E. (2011). The cognitive reflection test as a predictor of performance on heuristics-and-biases tasks. *Memory & Cognition, 39*, 1275–1289. https://doi.org/10.3758/s13421-011-0104-1

Trawalter, S., Richeson, J. A., & Shelton, J. N. (2009). Predicting behavior during interracial interactions: A stress and coping approach. *Personality and Social Psychology Review, 13*, 243–268. https://doi.org/10.1177/1088868309345850

Unzueta, M. M., & Lowery, B. S. (2008). Defining racism safely: The role of self-image maintenance on white Americans' conceptions of racism. *Journal of Experimental Social Psychology, 44*, 1491–1497. https://doi.org/10.1016/j.jesp.2008.07.011

Vitriol, J. A., & Moskowitz, G. B. (2021). Reducing defensive responding to implicit bias feedback: On the role of perceived moral threat and efficacy to change. *Journal of Experimental Social Psychology, 96*, Article 104165. https://doi.org/10.1016/j.jesp.2021.104165

Vorauer, J. D. (2006). An information search model of evaluative concerns in intergroup interaction. *Psychological Review, 113*, 862–886. https://doi.org/10.1037/0033-295X.113.4.862

Wetherell, M., & Potter, J. (1992). *Mapping the language of racism: Discourse and the legitimation of exploitation*. New York: Columbia University Press. New York.

Williams, M. T. (2020). Microaggressions: Clarification, evidence, and impact. *Perspectives on Psychological Science, 15*, 3–26. https://doi.org/10.1177/1745691619827499

Zhang, Z., Zyphur, M. J., & Preacher, K. J. (2009). Testing multilevel mediation using hierarchical linear models: Problems and solutions. *Organizational Research Methods, 12*, 695–719. https://doi.org/10.1177/1094428108327450