Justifying Dissent[*]

Leonardo Bursztyn[†] **Georgy Egorov**[‡] **Ingar Haaland**[§]
Aakaash Rao[¶] **Christopher Roth**[‖]

December 2022

Abstract

Dissent plays an important role in any society, but dissenters are often silenced through social sanctions. Beyond their persuasive effects, rationales providing arguments supporting dissenters' causes can increase the public expression of dissent by providing a "social cover" for voicing otherwise-stigmatized positions. Motivated by a simple theoretical framework, we experimentally show that liberals are more willing to post a Tweet opposing the movement to defund the police, are seen as less prejudiced, and face lower social sanctions when their Tweet implies they had first read credible scientific evidence supporting their position. Analogous experiments with conservatives demonstrate that the same mechanisms facilitate anti-immigrant expression. Our findings highlight both the power of rationales and their limitations in enabling dissent and shed light on phenomena such as social movements, political correctness, propaganda, and anti-minority behavior.

Keywords: Dissent; rationales; social image; social media
JEL Classification: D83, D91, P16, J15

[*] We thank the editors, Andrei Shleifer and Lawrence Katz, and five anonymous referees for very insightful comments. We also thank Davide Cantoni, Daniel Gottlieb, Ro'ee Levy, Pietro Ortoleva, Marco Tabellini, David Yang, Noam Yuchtman, and numerous seminar participants for helpful suggestions and Stelios Michalopoulos for a highly constructive discussion. Danil Fedchenko, Maximilian Fell, Takuma Habu, Hrishikesh Iyengar, Melisa Kurtis, and Stan Xie provided outstanding research assistance. We gratefully acknowledge financial support from the Pearson Institute for the Study and Resolution of Global Conflicts and the UChicago Social Sciences Research Center. Roth acknowledges funding from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy EXC 2126/1-390838866. The research described in this article was approved by the University of Chicago Social and Behavioral Sciences Institutional Review Board.

[†] University of Chicago and NBER, bursztyn@uchicago.edu

[‡] Kellogg School of Management and NBER, g-egorov@kellogg.northwestern.edu

[§] NHH Norwegian School of Economics, ingar.haaland@nhh.no

[¶] Harvard University, arao@g.harvard.edu

[‖] University of Cologne and CEPR, roth@wiso.uni-koeln.de

1 Introduction

From speaking out against injustice to victimizing protected groups, dissent can be a force for or against social change and therefore plays a consequential role in any society. Fundamental to dissent are *rationales* — narratives disseminated by political entrepreneurs, social movements, and media outlets — that provide arguments supporting dissenters' causes. Some rationales spur dissent through persuasion: they change people's views and, as a result, their public behavior. Yet dissent is often limited not because few people hold dissenting opinions, but rather because these people fear speaking their mind. Indeed, 62 percent of Americans agree that "The political climate these days prevents me from saying things I believe because others might find them offensive" (Ekins, 2020).

Consider Democrats who oppose the movement to defund the police. In many settings, publicly expressing this opposition generates social costs: opposition to police defunding may be seen as a signal of racial intolerance. Suppose that a credible study is publicized suggesting that defunding the police would increase violent crime. This new study might increase an individual's willingness to publicly oppose police defunding even if the study does not change her convictions, as long as she is able to *attribute* her views to the study. The key point is that the availability of this rationale opens up explanations other than racial intolerance for her position, reducing the social costs incurred by voicing it publicly and thus making her more willing to dissent.

In this paper, we explore the power and potential limitations of rationales in facilitating the expression of dissent. We present a simple theoretical framework demonstrating that rationales introduce "signal-jamming" that has important strategic consequences: by hindering the audience's ability to infer that a dissenter truly holds extreme beliefs, rationales lower the social cost of dissent and thereby increase the share of people willing to express their stigmatized beliefs publicly. Motivated by this framework, we experimentally examine the expression and interpretation of dissent in two contentious and policy-relevant domains: liberals' opposition to defunding the police and conservatives' support for deporting illegal immigrants. We focus on social media, where rationales from both mainstream and fringe sources are abundant and where people often face large social costs of expressing controversial opinions.

We begin by studying opposition to police reform among liberals. In a first experiment, respondents read a Washington Post article written by a Princeton criminologist arguing that "One of the most robust, most uncomfortable findings in criminology is that putting more officers on the street leads to less violent crime".[1] Respondents then choose whether to join a campaign opposing the movement to defund the police and, conditional on doing so, decide whether to post a Tweet promoting the campaign. The experimental manipulation subtly varies the availability of a social cover in the Tweet while holding fixed other potential motives to post. In particular, in the *Cover* condition, respondents' Tweets indicate that they were shown the article *before* joining the campaign, while in the *No Cover* condition, respondents' Tweets indicate that they were shown the rationale *after* joining the campaign.[2] The implied timing in the *Cover* condition provides these respondents with a social cover — the (implicit) justification that they joined the campaign because they were persuaded by the article's claims — while the timing implied by the *No Cover* condition eliminates this social cover. Differences in the "willingness to Tweet" thus cannot be explained by the persuasiveness of the rationale — all respondents in both groups read the article — or by respondents' expectations that the rationale will persuade their followers — both versions of the Tweet contain an identical description of and link to the article.

The availability of a social cover strongly affects posting behavior: in two preregistered waves of the experiment spaced a year apart, respondents are 11 percentage points more likely to post the Tweet in the *Cover* condition than in the *No Cover* condition. In two placebo experiments with an identical design, but with Tweets expressing support for causes associated with less stigma — as confirmed by an auxiliary survey — we find no difference between posting rates in the *Cover* and *No Cover* conditions. This evidence suggests that effects are indeed driven by (anticipated) changes in the stigma associated with dissenting expression rather than some other independent effect of the treatment. Several additional experiments provide further evidence for this interpretation and insight into the underlying mechanisms.

We conduct a second experiment to examine how the social cover shifts an audience's inferences about the motives underlying dissent and the resulting sanctions levied upon dissenters. Respondents are matched with a participant who posted the Tweet from the previous experiment — either a previous participant assigned to the *No Cover* condition or to the *Cover* condition — and are shown the anti-defunding Tweet their matched participant chose to post. They choose whether to deny a bonus to their matched participant,

[1] See "Why do we need the police?" Sharkey, Patrick. *The Washington Post*, June 12, 2020.

[2] Both Tweets are factually correct, as respondents in both conditions were shown the article both before and after joining the campaign.

a measure of social sanctions. We also elicit respondents' inferences about their matched participant's underlying prejudice: respondents guess whether or not the participant authorized a donation to a pro-Black organization.

The results confirm that the availability of social cover shifts inference and resulting social sanctions. Respondents matched with a participant in the *Cover* condition are 7 percentage points more likely to think that their matched participant authorized the pro-Black donation (relative to a *No Cover* mean of 27 percent) and are 7 percentage points less likely to deny their matched participant the $1 bonus (relative to a *No Cover* mean of 47 percent).

We next study the effects of rationales among a different sample, conservatives, and in a different policy context, anti-immigrant policies. Here, supporting the immediate deportation of all illegal immigrants from Mexico is a stigmatized opinion that people may be reluctant to publicly express, but a similar rationale as studied in the previous experiments — concerns about crime — may be effective in shifting inference about motives and thus decreasing social sanctions. In addition to speaking to the robustness of our previous findings and examining the use of rationales by a different population (conservative rather than liberal respondents), these experiments allow us to examine how rationales can generate social cover vis-a-vis different types of audiences. In particular, opposition to police defunding is primarily stigmatized by liberals' in-group (fellow liberals) rather than their out-group (conservative); in contrast, support for deportation is primarily stigmatized by conservatives' out-group (liberals) rather than their in-group (fellow conservatives).

The experimental manipulation follows the logic in our first experiment: in the *Cover* condition, respondents' Tweets indicate that they were exposed to a rationale — a clip of Fox News anchor Tucker Carlson arguing that illegal immigrants commit violent crimes at vastly higher rates than citizens — *before* joining the campaign, while in the *No Cover* condition, respondents' Tweets indicate that they were exposed to the rationale *after* joining the campaign. Our findings corroborate the importance of rationales in facilitating the expression of dissent: respondents are 17 percentage points more likely to post the Tweet in the *Cover* condition than the *No Cover* condition, relative to a *No Cover* mean of 47 percent. A further experiment shows that this rationale once again has strong effects on inference: respondents matched with a participant who chose to post the *Cover* Tweet are 5 percentage points more likely to believe that this participant authorized the pro-immigrant donation (relative to a *No Cover* mean of 9 percent) and are 7 percentage points less likely to deny their matched participant the bonus (relative to a *No Cover* mean of 80 percent).

Taken together, our evidence highlights the importance of rationales in facilitating dissent on both sides of the political spectrum; and it sheds light on the mechanisms by which individuals and institutions can influence public behavior by shaping the supply of rationales and perceptions of their social acceptability. Our findings have important implications for how the expression of dissent responds to the availability of new narratives. First, rationales are only effective to the extent to which observers believe that they genuinely change the dissenter's beliefs: an obscure or non-credible rationale may fail to shift inference, and may even backfire, if it signals the dissenter's underlying type. For example, if only intolerant people tend to read a particular source, citing a novel rationale provided by this source will fail to generate social cover. This implies that the endorsement of rationales by prominent figures such as politicians or celebrities may generate particularly large "social amplifiers": such figures may not only be more credible and *directly* persuade more people, but also more able to generate *common knowledge* such that dissenters can claim they were exposed to the rationale without seeking it out directly from stigmatized sources.

Conversely, groups seeking to suppress dissent have strong incentives to silence or marginalize potential sources of rationales (for example, disinviting campus speakers or branding certain news sources as fringe), because these tactics reduce the perceived probability that people will be exposed to rationales "by chance." If successful, these groups can create and sustain a "political correctness" culture — for better or for worse — in which certain rationales are ineffective because citing the stigmatized source undermines social cover. Indeed, at the time of our experiment, only 25% of Democrats privately supported decreasing police funding (Parker and Hurst, 2021). By challenging the credibility of rationales or explicitly linking them to stigmatized positions, a vocal group, even a vocal *minority*, can silence a majority.

Related Literature Our paper contributes to an emerging literature on narratives as drivers of economic and political behavior (Michalopoulos and Xue, 2021; Shiller, 2017). Related to our work is Foerster and van der Weele (2021), which studies the communication of rationales for and against donating to prosocial causes, and Bénabou et al. (2020), which models the production and circulation of justifications for morally questionable actions. Our contribution to this literature is to characterize and experimentally identify an important channel — the "social cover" effect — through which narratives, specifically rationales, shape the expression and the interpretation of dissent. Our theoretical framework and experimental evidence suggest means by which individuals and institutions can exploit this channel to facilitate or suppress dissent.

Therefore, our work also relates to a literature examining how social norms influence public behavior (Kuran, 1997; Bénabou and Tirole, 2006; Ali and Lin, 2013; Lacetera and Macis, 2010; Perez-Truglia and Cruces, 2017), and to a theoretical literature on political correctness (Morris, 2001; Golman, 2021). Like Braghieri (2022), we examine the role of social image concerns in shaping political correctness equilibria, though we investigate how rationales shape expression and interpretation rather than how differences between private and publicly-stated views lead to information loss. As in some of this previous work (Bursztyn et al., 2020a,b), our paper examines how previously-stigmatized public behavior can become socially acceptable, but a crucial conceptual difference is that our mechanism conditions social acceptability on the availability of a publicly observable rationale rather than the existence of misperceptions. This has important implications for interpretation and expression of dissenting views. In particular, rationales make public actions less informative about dissenters' underlying type and increase the public expression of dissent by lowering its social cost. This enables moderates who previously would have been unwilling to express dissent for fear of being labeled an extremist to voice their opinions, further hindering inference about dissenters' underlying type. In other words, our mechanism generates a "social amplifier" that magnifies rationales' persuasive effects.[3] We discuss how political entrepreneurs can strategically supply rationales to make the expression of unpopular views more mainstream.

This latter channel helps explain the mechanisms by which media and propaganda can promote socially undesirable behavior, such as anti-minority violence (e.g. Yanagizawa-Drott 2014; Adena et al. 2015; Enikolopov and Petrova 2015). Studies in this vein examining persuasion in field settings often find substantial effects (e.g. Caprettini et al. 2021) — in contrast to the relatively small effects of persuasion typically documented in a vast literature using information provision experiments (Haaland et al., 2021)). Among other plausible explanations for this discrepancy is the "social amplifier" channel: widespread propaganda creates common knowledge of rationales, generating greater social cover and

[3] In contrast to the information aggregation mechanisms examined in (Bursztyn et al., 2020a,b), rationales may facilitate the expression of views that are *privately* unpopular. Of course, the two mechanisms may be mutually reinforcing. For example, dissenting views may initially emerge among only a small segment of the population, which may employ rationales to lower the cost of publicly expressing these views to the rest of society. As a consequence of this public expression, others may then be privately persuaded. An information aggregation mechanism, such as an election, can then bring these previously-fringe views into the mainstream.

magnifying the effect of rationales on public behavior. Thus, our work also connects to a literature on populist political movements (e.g. Acemoglu et al. 2013; Guriev and Papaioannou 2020; Patir et al. 2021) insofar as authoritarian populists are often highly skilled at producing and disseminating rationales normalizing the victimization of minority groups.

Finally, our paper relates to a lab experimental literature documenting that individuals seize upon even flimsy excuses for selfish behavior.[4] Because behavior is typically private in these settings, these findings can be understood through a behavioral model of self-signaling, as in Bénabou and Tirole (2011) (similarly, Grossman and Van Der Weele 2017 formalize a mechanism by which individuals engage in willful ignorance as an excuse for selfish behavior). Our work holds this "self-excuse" channel constant — all individuals in our experiments privately voice their agreement with the Tweet — and we instead examine the role of rationales vis-a-vis *others*, shedding light on how rationales affect the expression, interpretation, and social punishment of dissent.[5] Our framework highlights levers by which agents can strategically manipulate the availability or credibility of rationales in order to influence dissent.

The remainder of this paper proceeds as follows. In Section 2, we present a simple model of the use and interpretation of rationales facilitating dissenting expression. In Section 3, we present experiments studying how the availability of a social cover shapes liberal respondents' willingness to publicly oppose the movement to defund the police, and how this social cover shifts their audience's beliefs about and behavior toward them. In Section 4, we present similar experiments focusing on conservative respondents in the context of anti-immigrant expression. Section 5 discusses implications of our findings and concludes. We list all main and auxiliary experiments in Appendix Table B.1.

2 Theoretical Framework

To organize these ideas and guide the experimental design, we start with a theoretical framework. All formal proofs are provided in Appendix A.

[4] See, for example, Dana et al. (2007); Hamman et al. (2010); Cunningham and de Quidt (2015); Lazear et al. (2012); Exley (2016); Golman et al. (2017); Saccardo and Serra-Garcia (2020) for work in economics and Shalvi et al. (2015) for a review of the extensive literature in psychology.

[5] A seminal contribution in psychology is Langer et al. (1978), which finds that individuals are more likely to comply with a request when it is justified by a reason, irrespective of whether the reason is good or bad. The authors interpret this as evidence for the "mindlessness of ostensibly thoughtful action", arguing that people have simply been conditioned to comply with requests accompanied by justifications.

2.1 Setup

The society N consists of a continuum of citizens facing a binary policy decision between the status quo (Q) and change (C). There is some objective measure of social welfare from decision C, and we denote this value w. The welfare under the status quo Q is normalized to zero. From the citizens' perspective, this value is distributed normally: $w \sim \mathcal{N}\left(w_0, \sigma_w^2\right)$. This social welfare may incorporate the expected economic payoff to each citizen from enacting decision C, but it may also include externalities to people outside the society or other factors inasmuch as citizens care about them.

Apart from the objective economic consequences captured by w, citizens have idiosyncratic tastes. Specifically, citizen i gets additional utility t_i if policy C, as opposed to Q, is enacted; we refer to t_i as i's type. We assume that t_i is distributed with c.d.f. $H(\cdot)$ and p.d.f. $h(\cdot)$, has finite mean $\mathbb{E}t = \bar{t}$, and satisfies the monotone hazard rate property.[6] To avoid corner cases, we assume that t_i has full support on the real line.

A citizen $i \in N$ is given a chance to publicly state support for change (decision $d_i = 1$) before an audience A. Doing so results in expressive benefit B and social cost S, so $U_i(d_i = 1) = B - S$.[7] We assume that

$$B = \beta\left(\mathbb{E}(w \mid *) + t_i\right);$$

in other words, the benefit is proportional to the sum of citizen i's posterior belief about w using all available information and i's own type. The social cost S is borne because action $d_i = 1$ may be revealing about i's type t_i, and having a high type is stigmatized by the audience.[8] For simplicity, we assume that stigma is linear in the audience's posterior about citizen i's type:

$$S = \gamma\left(\mathbb{E}_{-i}(t_i \mid d_i = 1, *) - \bar{t}\right).$$

In other words, a citizen pays a higher social cost if the audience's conditional expectation

[6] That is, $\frac{h(x)}{1-H(x)}$ is increasing in x, which is satisfied, e.g., for the normal and uniform distributions.

[7] By "expressive benefit," we mean utility derived from voicing one's true view independently of the social consequences. This might capture aversion to lying and/or staying silent on issues one cares about or other identity considerations.

[8] Note that by an audience, we do not necessarily mean the whole society, but rather the subset of individuals who pay attention to and judge the citizen for supporting the change. For example, a majority of citizens may support the change C, but if the people who listen and make inferences about the sender's type disproportionately support the status quo Q, or if the judgments of these people disproportionately matters to the citizen expressing support for C, the audience should be thought of as mainly consisting of Q-types.

of their type is higher than the unconditional one; this would be the case, for example, if the relevant audience that pays attention to i's statement and judges citizen i consists of supporters of status quo Q, or if their opinions matter to citizen i disproportionately. The utility from inaction ($d_i = 0$) is normalized to 0: $U_i\left(d_i = 0\right) = 0$.[9]

2.2 Analysis

In the absence of new information, the posterior of citizen i about w equals the prior w_0, and thus the benefit of action $d_i = 1$ is $B = \beta\left(w_0 + t_i\right)$. Citizen i makes the decision holding his social cost S fixed, and so chooses $d_i = 1$ if and only if

$$t_i \geq \frac{1}{\beta} S - w_0.$$

Thus, any equilibrium takes the threshold form, with the threshold τ_0 satisfying the condition

$$\tau_0 = \frac{\gamma}{\beta} \mathbb{E}\left(t_i \mid t_i > \tau_0\right) + k - w_0.$$

Generally speaking, the threshold need not be unique due to strategic complementarity: if not only extreme but also moderate types choose $d_i = 1$, the social cost is lower, which increases citizens' propensity to choose $d_i = 1$. However, if the distribution of t_i satisfies the monotone hazard rate property, the equilibrium is unique.

Proposition 1. *Suppose that $\gamma < \beta$. Then there is a unique equilibrium that takes the form of a threshold: individuals with $t_i > \tau_0$ choose $d_i = 1$ and those with $t_i < \tau_0$ choose $d_i = 0$.*

In other words, the equilibrium is unique provided that the citizen's choice is not driven solely by social image concerns and that the expressive benefit from their choice is sufficiently high.

Persuasive rationales Suppose that citizen i, prior to making the decision, receives an informative signal $s = w + \varepsilon$, where $\varepsilon \sim \mathcal{N}\left(0, \sigma_\varepsilon^2\right)$. Citizen i's posterior expectation of w

[9] We implicitly assume that the audience does not observe that i had a chance to make the action, and thus if he chooses $d_i = 0$ he is pooled with a continuum of citizens who are passive in this model. If the audience observes that inaction is by choice, there may be social consequences in this case as well. Nevertheless, all the results go through as stated.

is then equal to

$$w_1 = \mathbb{E}\left(w \mid s\right) = w_0 \frac{\sigma_\varepsilon^2}{\sigma_w^2 + \sigma_\varepsilon^2} + s \frac{\sigma_w^2}{\sigma_w^2 + \sigma_\varepsilon^2},$$

which exceeds w_0 if and only if $s > w_0$. If indeed the signal is positive ($s > w_0$), then for a fixed social cost S, this would prompt more citizens to choose $d_i = 1$ (specifically, all citizens with $t_i \geq \frac{1}{\beta}S - w_1$ would do so). This corresponds to a *persuasion* mechanism. In addition, if the audience is aware that more moderate people choose $d_i = 1$, the social cost of doing so is lower: intuitively, publicly supporting C is no longer a conclusive sign of extremism. Of course, a decrease in S will prompt even more people to choose $d_i = 1$, which corresponds to a *social amplifier* mechanism.

In practice, rationales trigger both persuasion and social amplifier mechanisms. This paper experimentally isolates the latter. To highlight the underlying theory, consider three cases. The equilibrium in each case takes a similar threshold form, but the thresholds themselves, and the social costs of dissenting, vary between cases. In the first case, the rationale is known neither to the sender nor to the audience: we refer to the associated equilibrium cutoff and equilibrium social cost as τ_0 and S_0, respectively. In the second case, the rationale is privately known to the sender, while the audience is unaware that the sender knew the rationale when making decision d_i: we denote the cutoff and social cost as τ_{priv} and S_{priv}, respectively. In the third case, the fact that the sender received the rationale is common knowledge: we denote the cutoff and social cost as τ_{pub} and S_{pub}, respectively. Intuitively, the difference between the first and second cases captures the effect of persuasion, while the difference between the second and third cases captures the role of the social amplifier mechanism. This is formalized in the following proposition.

Proposition 2. *Suppose that the informative signal satisfies $s > w_0$. Then a citizen who received this signal has a higher posterior about w than the prior. The equilibrium thresholds satisfy $\tau_0 > \tau_{priv} > \tau_{pub}$ and $S_0 = S_{priv} > S_{pub}$. Furthermore, an increase in σ_ε^2 weakens all these effects, and as $\gamma \to 0$, the differences between τ_{pub} and τ_{priv} and between S_{pub} and S_{priv} vanish.*

In other words, the ex-ante probability that citizen i chooses $d_i = 1$ is increasing from the case of no rationale to the private signal case to the public signal case, and the equilibrium social cost is the same in the first two cases, but decreases in the case of public signal. All these effects are attenuated if the signal is noisier and therefore less informative: citizens update less and are less likely to choose $d_i = 1$, and the associated social cost does not increase as much either. Practically, this means that if the same information is obtained from a less credible source, the changes in behavior and social cost will be smaller, and in the limit, an uninformative signal will have no effect. Lastly, in the absence of social image concerns, the social amplifier effect disappears: that is, we should observe no difference in behavior between the public and private signal cases in non-stigmatized contexts, though there may still be a persuasion effect.

2.3 Polarizing Rationales

In reality, individuals are often presented with the same evidence, but the evidence has heterogeneous consequences (e.g. some individuals react favorably to news that a neighborhood is diversifying, while others react unfavorably) or is interpreted differently (e.g. due to differences in background knowledge, cognitive limitations, or behavioral biases). Can rationales still be effective even if they are not persuasive *on average* — that is, they "dissuade" as many people as they persuade? In Appendix Section A.3, we show that they can. The intuition is that as long as the rationale changes some people's views, the audience faces an inference problem. Assuming for simplicity that citizen i may either get a high signal $s_h > w_0$ or low signal $s_l < w_0$, the audience knows that the citizen i who chose $d_i = 1$ may have done so either because t_i is high, or because i got a high signal s_h. More precisely, the set of citizens who would choose to support change C now contains some types with $t_i < \tau_0$ (moderates who got a high signal s_h) and lacks some types with $t_i > \tau_0$ (extremists who got a low signal $s_l < w_0$). As long as the share of the former is not too small, the posterior of t_i conditional on choosing $d_i = 1$ goes down. As a result, more citizens will choose $d_i = 1$ and will face a lower social cost from doing so. Put differently, for a rationale to be effective it does not have to be persuasive on average, so long as it hinders inference about the motives underlying the stigmatized action.

3 Opposition to Defunding the Police

The experiments in this paper examine the expression of dissent on social media. Expression on social media is of direct interest: over 70 percent of Americans report using social media daily, many politicians and other prominent figures have turned to social media as a primary channel of communication with the public, and social media has been linked to a number of important real-world outcomes: protests (Enikolopov et al., 2020), hate crimes (Müller and Schwarz, 2018; Bursztyn et al., 2019), and social movements (Levy and Mattsson, 2021). Second, expressing dissent on social media — like doing so in real-world offline settings, and unlike doing so in more artificial lab settings — may have real social costs vis-a-vis a natural population about whose opinions respondents care — family members, friends, acquaintances, and current and/or future employers. Indeed, a substantial majority of hiring managers report using social media accounts as a screening tool (O'Brien, 2018).

Our first two experiments examine the use and interpretation of rationales for opposing the movement to defund the police. The slogan "defund the police" rose to national prominence after the murder of George Floyd in May 2020; advocates seek to decrease funding for police departments, and many favor restricting the responsibilities of law enforcement primarily to violent crime, redirecting resources to specialized response teams such as social workers and conflict-resolution specialists to deliver other services (Thompson, 2020). Popular opposition to police defunding is relatively high: as of an October 2021 Pew Research survey, only 15 percent of adults, 25 percent of Democrats, and 23 percent of Blacks support reducing spending on policing in their area (Parker and Hurst, 2021). Nonetheless, because the movement is closely linked to concerns about racial injustice — most advocates claim that the American law enforcement system is fundamentally racist and requires radical reform (or abolition) — it seems *a priori* plausible that many liberals would feel uncomfortable publicly voicing opposition to defunding. This is particularly true given that liberal Twitter users are more interested in social justice causes and are more likely to call out perceived injustice than liberals at large (Cohn and Quealy, 2019). Indeed, in a pre-registered survey (Auxiliary Survey 1), we find that 80% of Democrats anticipate "Strong social backlash" or "Significant social backlash" if they were to express opposition to police defunding on social media.[10]

3.1 Experiment 1: Rationales and Anti-Defunding Expression

3.1.1 Motivation for experimental design

Experiment 1 studies how the social cover provided by rationales affects respondents' willingness to post a Tweet on their account opposing the movement to defund the police. Identifying this effect is challenging from both a design and ethical perspective. From a design perspective, we need to manipulate the availability of a social cover, ruling out

[10] The pre-registration can be accessed at `https://aspredicted.org/7nm5j.pdf`. See Appendix E.5 for experimental instructions.

other possible reasons for why a rationale might change posting behavior. For example, the rationale may affect posting behavior by changing respondents' private beliefs (persuasion), or respondents might cite the rationale to persuade others (anticipated persuasion). Identifying the cover effect requires us to hold these other channels fixed across experimental conditions. At the same time, we wish to avoid a complicated or heavy-handed intervention in order to maximize the extent to which our results can speak to the expression of dissent in real-world contexts. From an ethical perspective, while we want to examine the most natural possible outcome — respondents' willingness to Tweet — we prefer to avoid leading respondents to actually post political content on Twitter (a particular concern in our similarly-structured Experiment 3, which studies willingness to publicly support a campaign to deport all illegal Mexican immigrants). A related and conflicting goal is to avoid explicitly deceiving respondents. We address these design and ethical difficulties with an experiment aiming to (1) hold the *persuasion* and *anticipated persuasion* effects constant while varying only the availability of a social cover, (2) measure respondents' revealed-preference willingness to express dissent on their Twitter account, (3) avoid having respondents actually posting these Tweets, and (4) avoid explicit deception.

3.1.2 Sample and experimental design

We conducted our pre-registered Experiment 1 in October 2021 with a sample of 1,122 Democrats and Independents.[11] As explained below, this resulted in a final sample for analysis of 523 respondents. We then conducted a pre-registered replication of the experiment (Experiment 1R) in October 2022 targeting the same final sample size.[12] For both Experiment 1 and Experiment 1R, we recruited respondents from both Luc.id and CloudResearch, two survey providers widely used in the social sciences (Litman et al., 2017; Wood and Porter, 2019).

Figure 1 outlines the structure of Experiment 1. After completing a short attention check, we ask respondents to log in to our survey using their Twitter account through "Tweetability," a Twitter application we created using Twitter's Application Programming Interface (API) that allows us to schedule Tweets to be posted on the users' accounts at a future date. To an observer, these Tweets look as though they were posted by the

[11] See Appendix Table B.1 for all pre-registration IDs. The full set of experimental instructions is included in Appendix E.1.

[12] Due to changes in the sampling interface of our survey provider, we targeted only Democrats in Experiment 1R. The experimental instructions are identical to those in the original experiment with the exception of additional post-treatment questions, as discussed below.

respondent him or herself. We automatically capture respondents' Twitter handles after they log in. Respondents are assured that we will never use this application to access any private information from accounts, that all data will be securely stored until its deletion by no later than December 1, 2021 (2022 for Experiment 1R) and that we will never schedule posts on their accounts without their explicit permission. Respondents then respond to a set of basic demographic and other background questions.

We then present respondents with an op-ed written in the Washington Post by Patrick Sharkey, a professor of public affairs and criminology at Princeton University.[13] In the article, Sharkey argues that a vast body of evidence shows that increasing policing decreases violent crime, that defunding the police is thus likely to increase violence, and that other solutions (e.g. granting communities more resources to maintain safety) will likely be more effective. After reading the article, respondents are asked if they would like to join a campaign to oppose the movement to defund the police. The survey terminates for respondents who do not join. Respondents who join are presented with the article again and informed that they can spend as long as they wish reading it.

Once they continue, we inform respondents that the campaign involves circulating a petition on Twitter opposing the movement to defund the police. We show them a screenshot of the Tweet and ask if they are willing to schedule the Tweet to be posted on their account. We inform respondents that the Tweets of all respondents will be posted if and when we have surveyed people in all US counties (a strategy which, as we explain to respondents, is often used in social media campaigns to make certain topics "trend" on users' timelines). In practice, because we target fewer respondents than the number of counties in the US, we ensure Tweets will never be posted. Our outcome can nonetheless be interpreted as revealed-preference conditional on respondents believing it sufficiently probable that we will reach respondents in all counties.[14]

Respondents in the *Cover* condition are asked whether they would like to schedule the following Tweet:

> I have joined a campaign to oppose defunding the police: [LINK]. Before joining, I was shown this article written by a Princeton professor on the strong scientific

[13] The article is available at `https://www.washingtonpost.com/outlook/2020/06/12/defund-police-violent-crime/`.

[14] It is possible that some respondents believe it unlikely that the Tweets will be posted, but for this to bias our estimated treatment effects, we would require not only that this belief is differential across treatment conditions but also that respondents who hold this belief are more or less likely to authorize the *Cover* Tweet relative to the *No Cover* Tweet.

> evidence that defunding the police would increase violent crime: [LINK]

The Tweet is identical for respondents in the *No Cover* condition, with one exception: the second sentence begins "**After** I joined the campaign…". Both Tweets are factually correct (all respondents were in fact shown the article both before and after joining the campaign), but this difference in wording suggests to potential readers of the Tweet that respondents in the *Cover* condition had been exposed to the scientific evidence against defunding the police before joining the campaign — and thus had a strong rationale for doing so. In contrast, the *No Cover* Tweet suggests that respondents had only been exposed to the evidence after joining, and thus that the evidence could not have led them to join the campaign. This design therefore isolates the cover effect of rationales while fixing the persuasion channel (all respondents are exposed to the same information) and the anticipated persuasion channel (all respondents know their Tweet's readers will see a link to the article in the Tweet, conditional on the Tweet being posted) across conditions.[15] By employing a one-word manipulation, we also hold other potential confounds, such as the length of the Tweet, fixed across conditions. Our final sample is well-balanced on observables across treatment arms (Appendix Table B.2).

Discussion of ethical considerations Although our experiment avoids explicit deception — all statements subjects see are factually true — our design clearly misleads subjects: they believe that their Tweets might be posted (if we recruit respondents in every US county), when in fact we purposefully recruit fewer respondents than the number of counties such that there is no chance this condition will ever be met. In experimental economics, deceiving or misleading respondents is often considered problematic due to concerns that it will lead subjects to expect deception in future experiments, potentially changing their behavior. Because subjects do not know, and never learn, that we recruited fewer respondents than the number of US counties, this concern does not apply to our experiment. More generally, we concluded that the benefits of protecting participants' privacy and avoiding contributing to a political campaign outweighed the costs of misleading

[15] One potential confound, which we cannot fully rule out, is that the respondent updates negatively about the utility they will derive from joining due to anticipated social interactions with other people who joined the campaign. While there are no differences between treatment conditions until and including the screen when respondents choose whether or not to join the campaign, and thus respondents should have identical beliefs about who joins the campaign, they may particularly care about social interactions with others who post the Tweet, not just those who join the campaign. In practice, this is unlikely to significantly bias estimates: as described to respondents in the experimental instructions, the campaign revolves around posting a Tweet to one's followers, rather than interacting with other Twitter users who posted the Tweet.

respondents. Moreover, our design ensures that the Twitter *followers* of the respondents in our survey will not be misled by respondents' Tweets as to whether they read the article before or after joining the campaign — given that these Tweets are never posted. We discuss the ethical considerations underlying all experimental designs in greater detail in Appendix C.

3.1.3 Results

Table 1 displays the results separately for the main experiment and the replication. The results are similar in both waves, so we pool the two in the discussion below and in the leftmost comparison in Figure 2. 55% of respondents authorize the Tweet in the *No Cover* condition compared to 66% of respondents in the *Cover* condition ($p < 0.01$). These effects are stable to the inclusion of controls; the effect size corresponds to 0.25 standard deviations, comparable to or larger than the effects on persuasion generally documented in information provision experiments (Haaland et al., 2021) and the effects of image concerns generally documented in experiments varying the observability of decisions (Bursztyn and Jensen, 2015).[16] This relatively large effect underscores the importance of the cover effect in driving the expression of dissent.

Heterogeneity and external validity We can estimate treatment effects only for respondents who were willing to log in via our app and join the campaign. We provide experimental evidence that this selection is not driving our effects in Auxiliary Experiment 5, reported below, but we also shed light on the magnitude of potential selection by investigating treatment effect heterogeneity. In Column 1 of Appendix Table B.4, we show that there is muted treatment effect heterogeneity by age, race and ethnicity, gender, and education; as shown in Appendix Table B.5, our estimated treatment effects remain stable when we reweight the sample to match the general population on these observables.[17]

[16] Indeed, in our pre-registered Auxiliary Experiment 1 with the same rationale, we estimate a persuasion effect on private attitudes of 0.12 standard deviations (p=0.059). See Appendix B.1.3 for details, Appendix E.6 for experimental instructions, and Appendix D for balance and representativeness tables for all auxiliary experiments.

[17] In Experiment 1R, we collected additional information on the characteristics of respondents' Twitter accounts. In Appendix Table B.6, we show that treatment effects do not vary significantly by respondents' number of followers. There is some suggestive treatment effect heterogeneity by self-reported perception of the share of followers who would support defunding the police: treatment effects are driven by respondents who perceive this fraction to be between 30% and 70% of their followers. One way to interpret this finding is that respondents whose followers mostly disapprove of defunding the police may not need a cover, while those whose followers mostly approve may not be elastic to social cover given that they still expect

3.2 Ruling out alternative explanations

In this section, we consider alternative explanations for the treatment effects presented above.

3.2.1 Direct evidence on perceptions of differential misleadingness

To make our instructions as natural as possible, we present a plausible rationale for showing respondents the article again after they join the campaign. In particular, we write "Since you chose to join the campaign, we wanted to give you more time reading the [article]": a natural offer to someone who had expressed particular interest in the topic. Even so, one potential concern is that respondents are more willing to schedule the *Cover* Tweet ("Before joining the campaign...") than the *No Cover* Tweet ("After joining the campaign...") because they think the latter Tweet misleads respondents as to when they joined the campaign relative to reading the article.

Our first piece of evidence that this confound is not driving our treatment effects comes from two post-treatment questions we added to Experiment 1R. First, we ask respondents whether they perceived the Tweet to be misleading. Second, for those who answer that they did, we ask them to explain why they felt this was the case (in open-ended format), and we hand-code the responses.

Only 2 percent of respondents perceive the Tweet to be misleading. As shown in Panel A of Table 2, this fraction is in fact 2 percentage points higher in the *No Cover* group than in the *Cover* group, though the difference is not statistically significant. Of the respondents who indicate the Tweet was misleading, none write anything related to the timing of the information provision, the timing of joining the campaign, or the "before"/"after" wording (the latter being the only difference between treatments). Moreover, restricting the sample to respondents who indicate the Tweet is not misleading leaves treatment effects virtually unchanged.

We now turn to a series of experiments designed to provide further evidence against this and other potential confounds and to shed light on the underlying mechanisms. We summarize the results of these experiments in Table 2.

substantial social punishment. Finally, in Panel A of Appendix Table B.7, we show treatment effects by partisan affiliation. Overall, while there is some evidence of heterogeneity, we are generally underpowered for these comparisons. As shown in Column 2, our main treatment effects in Experiment 1 and Experiment 3 are robust to limiting the sample to Democrats and Republicans, respectively.

3.2.2 Placebo experiments

There may be reasons unrelated to the difference in perceived social cover that respondents prefer the *Cover* Tweet to the *No Cover* Tweet. Respondents may, for example, find the "After" wording strange or unnatural. To rule out that our estimates are a mechanical effect of the "Before"/"After" wording, we conduct two "placebo experiments" (Auxiliary Experiments 2 and 8), whereby "placebo experiments," we mean that the experiments replicate the manipulation of Experiment 1, but do so in less controversial domains in which, if the underlying mechanism driving our findings in Experiment 1 is indeed social cover, we would expect no treatment differences. One of the placebo experiments is in a relatively political domain, but with an uncontroversial policy where social sanctions are unlikely to exist: support for the conservation of the Amazon rainforest. The other placebo experiment is in a (relatively) apolitical domain where social sanctions are again unlikely to exist: eliminating daylight saving time.[18]

To confirm that social sanctions in either placebo domain are indeed less relevant, we return to the results of Auxiliary Survey 1, in which we ask respondents whether or not they privately support each of four causes: defunding the police (as in Experiment 1), conserving the Amazon rainforest, eliminating daylight saving time, and immediately deporting all illegal Mexican immigrants (as in Experiment 3). For those who privately support each cause, we ask whether they anticipate that they would face social backlash if they were to express this support on social media. Appendix Figure B.1 confirms that respondents who privately support defunding or deportation expect substantial backlash if they were to express their views on social media (59% and 71%, respectively, expect "significant" or "strong" backlash), while respondents who privately support rainforest conservation or eliminating daylight saving expect far less backlash for expressing these views on social media (20% and 18%, respectively).

Having confirmed that anticipated social backlash is far lower in the rainforest and daylight saving contexts, we turn to the design and manipulation of the placebo experiments, which are identical to Experiment 1 except for the settings and choice of rationales. For the Amazon experiment, the rationale is a Reuters article reporting a new study that finds that over 10,000 species are at risk due to deforestation in the Amazon; for the daylight saving experiment, the rationale is an article written by a Vanderbilt neurologist on the health costs of daylight saving time.[19]

[18] See Appendices E.7 and E.8 for experimental instructions.

[19] The Amazon Tweets read: "I've joined a campaign to immediately stop the destruction of the Amazon

Panels B and C of Table 2 show no significant difference between posting rates in the *Cover* and *No Cover* conditions for either experiment. Pooling the two placebos in the rightmost comparison of Figure 2, we estimate a tight null effect of *Cover* on posting rates. The large and significant difference in effect sizes between the defunding experiments and the placebo experiments suggest that effects are indeed driven by (anticipated) changes in the stigma associated with dissenting expression rather than some other independent effect of the before/after wording.[20]

Ultimately, however, there may be factors specific to the Amazon and daylight saving contexts, or the rationales we use, that lead to the lack of treatment effects of the *Cover* condition. Thus, while highly suggestive, our placebo results cannot definitely prove our preferred interpretation of the results in Experiment 1. For further evidence for this interpretation, and for evidence on the underlying mechanisms, we turn to a series of auxiliary experiments.

3.2.3 Addressing anticipated persuasion

It remains a possibility that respondents anticipate that the *Cover* Tweet will be more persuasive to followers than the *No Cover* Tweet, and that this difference drives our estimated treatment effects. Relatedly, it could be the case that respondents believe that their followers are more likely to read the article after seeing the *Cover* Tweet than after seeing the *No Cover* Tweet.

To mitigate concerns related to such differential anticipated persuasion, we run an auxiliary experiment (Auxiliary Experiment 3). In this experiment, we present Democratic and Independent Twitter users with either the *Cover* or *No Cover* Tweet and then ask them to estimate the share of their followers who would join the campaign after seeing

rainforest! [Before/After] I joined the campaign, I was shown this article about how 10,000 species risk extinction in the Amazon: LINK. Join the campaign and sign the petition: LINK". The daylight saving Tweets read: "I have joined a campaign to eliminate daylight saving time: LINK. [Before/After] joining the campaign, I was shown this article by a Vanderbilt professor of neurology on how daylight saving time is connected with serious negative health effects: LINK.".

[20] We cannot definitively rule out the possibility that the lack of treatment effects in either placebo is due to the sum of two countervailing effects: the social cover mechanism and another mechanism by which people prefer the "After" wording because it signals that they did not have to be informed about the issue to support it. While this confound could plausibly be present in the Amazon context, where people might want to signal that they are a "good" type who does not need to be persuaded in order to support rainforest preservation, we view it as much less likely in the daylight saving context, in which such signaling motives are implausible.

their Tweet, a summary statistic for the combined effects of all channels above.[21] Panel D of Table 2 shows a small and insignificant 1.9 percentage point difference; we can rule out differences of greater than 4.2 percentage points with 95% confidence. This suggests that differences in posting rates are not driven by differences in the anticipated persuasiveness of the Tweets, as respondents' posting decisions would need to be unrealistically elastic to their beliefs about their audience's persuadability in order to generate the 12 percentage point treatment effect documented in Experiment 1. We provide further evidence against this mechanism below.

3.2.4 Direct evidence on social cover mechanism

We now provide direct evidence that our manipulation varies the perceived availability of social cover, and that this availability is an important consideration on respondents' minds when considering the expression of dissent. We conduct Auxiliary Experiment 4 with a sample of 402 Democrats with Twitter accounts recruited from Prolific. This broader sample allows us to probe the external validity of our findings. In particular, respondents are not required to grant our "Tweetability" app permissions to schedule posts on their Twitter account, which may induce selection into Experiment 1.

Experimental design Respondents begin by reading the article presented in Experiment 1 describing the evidence that defunding the police would increase violent crime. We ask them to imagine that at this stage, they joined a campaign to oppose defunding the police. As in the main experiment, all respondents are then given the chance to read the article again.[22] Then, respondents randomized into the *Cover* condition are asked which of two Tweets they would *hypothetically* prefer to post: the Tweet from the *Cover* condition in Experiment 1, or a *Control* Tweet omitting any reference to a rationale:

> I have joined a campaign to oppose defunding the police: [LINK].

Respondents randomized into the *No Cover* condition are instead asked about their hypothetical preference between posting the Tweet from the *No Cover* condition in Experiment 1 or the *Control* Tweet above. After respondents choose their preferred Tweet, we ask them to "Please explain why you chose this Tweet rather than the other Tweet." Our object of interest is the difference in respondents' explanations between conditions.

[21] See Appendix B.1.4 for details and Appendix E.12 for experimental instructions.

[22] See Appendix E.13 for experimental instructions.

A few comments about the experimental design are in order. First, we separately study preferences for the *Cover* Tweet over the *Control* Tweet and for the *No Cover* Tweet over the *Control* Tweet, rather than directly estimating preferences for the *Cover* Tweet over the *No Cover* Tweet. Our design thus avoids making the "Before/After" distinction between the Tweets salient, better capturing behavior both in our main experiment and in real-world settings and reducing the scope for experimenter demand effects. Similarly, our use of open-ended text to elicit motives, rather than structured questions, avoids priming respondents on particular motivations and better captures what naturally comes to mind when making their choice.

We hand-code open-ended responses across three categories. "Social cover" responses mention that the respondent's preferred Tweet indicates to followers that the article affected the respondent's choice to join the campaign.[23] "Anticipated persuasion" responses mention that the article might persuade others.[24] Finally, "Information" responses mention that the article is informative or credible, or that it provides an explanation for why people might want to join the campaign, but do not explicitly relate the information to the respondent's own views or other people's views.[25] Many respondents classified as "Information" may have had the "Social cover" or "Anticipated persuasion" mechanisms in mind, but wrote responses that we could not unambiguously classify into either category. We chose a conservative coding scheme for "Social cover" and "Anticipated persuasion" in order to provide a plausible lower bound.

Results We begin by analyzing respondents' preferences over which Tweet to post. 83% of respondents in the *No Cover* condition prefer the Tweet linking to the evidence over the *Control* Tweet without the evidence, compared to 87% of respondents in the *Cover* condition.[26] The high fraction choosing the Tweet with the rationale (whether the *Cover*

[23] For example, one respondent writes: "I think the evidence provided in the article is an important catalyst in why I would have joined the campaign and without any context that first tweet could be misconstrued, or even cause me to be publicly shamed."

[24] For example, one respondent writes: "The tweet is meant to not only inform people of your decision, but to also advertise others to do the same. Having supporting evidence for your cause will increase the chance of others to side and agree with you. Tweet B does this, Tweet A doesn't."

[25] For example, one respondent writes: "I would want others to see this article and know that I have some evidence to back my tweet."

[26] The treatment effect is not comparable with the effect estimated in Experiment 1: for example, we might observe zero treatment effect in this experiment and a strong treatment effect in Experiment 1 if most respondents prefer the *Cover* Tweet to the *No Cover* Tweet, but strongly prefer either Tweet to the *Control* Tweet (while a minority of respondents exhibit strong preferences for the shorter *Control* Tweet). Nonetheless, it is reassuring that the treatment effect is positive (though statistically insignificant,

or the *No Cover* version) over the *Control* Tweet suggests a widespread preference for citing evidence when engaging in dissenting expression, while the high fraction choosing the *No Cover* version constitutes further evidence that respondents do not avoid the "After" wording due to concerns about it being misleading or unnatural.

We next turn to the open-ended text. The perceived social costs of dissent in this setting are further evidenced by the substantial number of responses mentioning some form of social sanctions. A relatively large fraction of respondents (20 percent) explicitly mention the social cover mechanism, three times the number who mention the anticipated persuasion mechanism (7 percent). The majority of responses (53 percent) fall into the "Information" category, though many responses in this category likely meant to convey concerns relating to social cover. Focusing on treatment effects across conditions, reported in Panel E.1 of Table 2, the one-word manipulation indeed induces substantially more respondents to mention social cover (a 10 percentage point difference, or a 67 percent effect relative to the *No Cover* mean).

Consistent with the results of Auxiliary Experiment 4, the manipulation appears to have no effect on the probability that respondents mention that their followers will find the article persuasive. While these two pieces of evidence cannot definitively rule out differences in the anticipated persuasiveness in the Tweet, they do suggest that any such differences are unlikely to drive the large treatment effects of the *Cover* condition that we document.

To gauge other potential confounds, we also hand-code responses along further dimensions. "Unnatural" responses mention that one Tweet seems more unnatural or strangely worded than another; "Misleading" responses mention that one Tweet seems more misleading or deceptive than another; "Signaling" responses mention that one Tweet suggests that the respondent supports the cause more strongly than the other; "Experimenter demand" responses mention that the experimenter wants the respondent to choose one Tweet over another, or that the respondents' followers will believe this is the case. As shown in Panel E.2 of Table 2, almost no Tweets fall into any of these categories.[27]

Together, the placebo experiments, the anticipated persuasion experiment, and this experiment eliciting participants' reasoning strongly suggest that the treatment effects documented in Experiment 1 are indeed driven by differences in the availability of a social

$p = 0.311$).

[27] Of the 15% of respondents who choose the *Control* Tweet without a rationale, two-thirds cite its shorter length as the reason for doing so. Given that the one-word manipulation in Experiment 1 holds the length of the Tweet fixed, preferences for shorter or longer Tweets will not affect our results.

cover.

3.2.5 The role of credibility

In Section 2, we show that the credibility of rationales matters: a rationale that is perceived to come from a questionable source, or whose credibility is otherwise undermined, is likely to be less effective.[28] The wording of the Tweet in Experiment 1 emphasizes the credibility of the rationale, explicitly stating that the author is a Princeton professor and that the article is based on strong scientific evidence; our theory implies that reducing the credibility of the rationale will reduce its effect on posting behavior and increase the associated social sanctions.

We examine the role of credibility with Auxiliary Experiment 5, which we also use to probe another dimension of external validity. In particular, the sample of Experiment 1 consists of respondents who were willing to grant our app permissions to post on their Twitter account, and thus is likely unrepresentative of the population of social media users.[29] To assess the importance of social cover in facilitating dissent among this broader population, we do not ask respondents in Auxiliary Experiment 3 to log in via Twitter; we instead ask whether respondents would have (hypothetically) been willing to post the Tweet.

Experimental design The design of Auxiliary Experiment 5 is almost identical to the design of Experiment 1.[30] All respondents who report actively using Facebook and Twitter are eligible to participate. As in Experiment 1, they read the Sharkey article and are given the opportunity to join the campaign to oppose defunding the police; those who do not join are screened out of the survey. Remaining respondents are presented the article a second time. We then explain to them that we are interested in whether they would be willing to make the post in question (either the *Cover* or the *No Cover* post) if it were included as a campaign feature. To probe mechanisms, we also ask an incentivized (post-outcome)

[28] In particular, it is not necessary that the audience finds the rationale persuasive, but rather that the audience thinks it is plausible that the *dissenter* him or herself was persuaded.

[29] To speak to the extent of selection by social desirability into "Tweetability" login, we follow Dhar et al. (2022), who use a 13-item version of the Marlowe-Crowne social desirability scale to measure respondents' concern for social approval (Crowne and Marlowe, 1960; Reynolds, 1982). We implement this scale in Auxiliary Experiment 3. Appendix Figure B.2 shows economically and statistically insignificant differences in this score between our experimental sample (which authorized the login) and the general population, suggesting that our sample is not selected on concerns for social approval.

[30] See Appendix E.14 for experimental instructions.

question eliciting perceived social sanctions: respondents estimate the share of Democrats who, upon seeing the post, chose to deny the poster a bonus. Finally, and most importantly, we cross-randomize a "credibility" manipulation with our previous manipulation of social cover, resulting in four conditions. In particular, to construct "lower-credibility" versions of the Tweets, we remove the references to Sharkey's academic credentials and to the scientific evidence underlying the article's claims. The revised lower credibility Tweets read:

> I have joined a campaign to oppose defunding the police: [LINK]. [Before/After] joining, I was shown this article arguing that defunding the police would increase violent crime: [LINK]

Our framework predicts that this less credible rationale will generate less social cover and thus will be less effective in facilitating dissent.

Results We present results in Panel F of Table 2. Restricting attention to the higher-credibility version of the post (i.e. the version used in Experiment 1), we find an almost identical treatment effect to that documented in Experiment 1, confirming that our results generalize to the broader sample of social media users. Turning to the lower-credibility version, we find a smaller and statistically insignificant treatment effect. While the results are qualitatively consistent with the predictions of Proposition 2, the difference between the high and low credibility condition is not statistically significant. Since we are generally not powered to detect significant interaction effects, we view the smaller effect size of the low credibility condition as suggestive evidence consistent with our theory.

We find a similar pattern when we instead examine respondents' guesses as to the number of Democrats who would deny a person who made the post a bonus (our measure of perceived social sanctions): respondents believe that the social cover is effective in reducing social sanctions when the rationale is highly credible. When the rationale is less credible, the effects on perceived social sanctions are smaller and statistically insignificant (though again we lack the statistical power to detect significant interaction effects).

The perceived treatment effect of the *Cover* condition on social punishment (relative to the *No Cover* condition) implied by our data is 5 percentage points in the high-credibility condition and 2 percentage points in the low-credibility condition. As we show in the next section (Experiment 2 and Auxiliary Experiment 6), the actual treatment effect of *Cover* on social punishment is 7 percentage points in the high-credibility condition and 1 percentage point in the low-credibility condition. In other words, respondents are well-calibrated about the treatment effects of *Cover*. They are also fairly well-calibrated about the *levels* of punishment: pooling across all conditions, they expect around half of Democrats to deny the bonus, relative to the actual share of 43%. Thus, our mechanism does not require respondents to over- or under-estimate the share of their audience who would sanction them for expressing dissent, nor does it require this share to be a substantial majority.

Discussion The manipulation arguably generates a fairly modest reduction in credibility (as it still features an article from The Washington Post, a well-respected outlet among liberals): far more modest than, for example, citing a right-leaning outlet or making such a claim without any supporting evidence. Nonetheless, even this modest reduction in credibility halves the estimated effect of the rationale on posting. While drawing general conclusions about credibility would require substantially greater evidence than we provide here, our evidence suggests that one way a vocal minority might silence public dissent is by setting the "credibility bar" high, accepting only overwhelmingly conclusive evidence as legitimate.[31] A society that sets this "credibility bar" too high may stifle the expression of legitimate perspectives on issues where strong evidence does not exist. Indeed, if the credibility bar *varies* between groups — for example, if conservatives are seen as more easily persuaded by fake news than liberals — then groups held to a lower credibility bar can use a wider variety of rationales and thus may be willing to dissent in a wider variety of contexts.

3.3 Experiment 2: Interpretation of Anti-Defunding Rationale

Our theoretical framework implies that rationales lower the social cost of dissent by making the action less informative about type. As documented in Section 3.1, respondents are more willing to dissent when they can draw upon credible rationales because they *expect* such rationales to reduce the informativeness of dissent for prejudice and thus lower the associated social costs. In Experiment 2, we examine whether rationales indeed serve this purpose.

[31] Only 25% of Democrats privately support decreasing funding for police in their area, compared with 34% of Democrats who privately support increasing funding (Parker and Hurst, 2021). Thus, the results of Experiment 1 and Auxiliary Experiment 5 jointly illustrate how public dissent can be silenced by a vocal minority. Through the lens of our theoretical framework, different audience members may contribute differently to overall social sanctions S: opponents of defunding may not sanction respondents who hold either opinion, while a significant fraction of supporters may heavily sanction opponents.

3.3.1 Sample and experimental design

We conducted our pre-registered Experiment 2 in November 2021 with a sample of Democrats and Independents recruited from Prolific.[32] Our final sample of 1,040 Democrats and Independents is mostly balanced on observables across treatment arms (Appendix Table B.9).

Figure B.3 outlines the structure of Experiment 2. After completing a battery of demographic and other background questions, respondents are informed that they have been matched with a previous survey participant who joined a campaign to oppose the movement to defund the police. They are then randomized into a *Cover* and a *No Cover* condition: respondents in the *Cover* condition are told that their matched participant authorized the Tweet corresponding to the *Cover* condition of Experiment 1 ("Before I joined the campaign. . . ") whereas respondents in the *No Cover* condition are told that their matched participant authorized the *No Cover* Tweet ("After I joined the campaign. . . ").

We begin by asking respondents the following open-ended question: "Why do you think your matched participant chose to join the campaign to oppose defunding the police?" This approach avoids priming respondents to think about particular dimensions and instead directly elicits "what comes to mind" (Gennaioli and Shleifer, 2010). As a more direct measure of inference about their matched participant's prejudice, we subsequently tell them that their matched participant had the opportunity to authorize a \$5 donation to the National Association for the Advancement of Colored People (NAACP) and ask them to guess whether or not the participant donated. Finally, we also give respondents the opportunity to authorize a \$1 bonus to their matched respondent (at no cost to themselves): declining to do so is our measure of social sanction.

3.3.2 Results

We estimate statistically and economically significant treatment effects on all three measures of inference. The leftmost comparison in Panel A of Figure 3 displays the fraction of participants in the *Cover* and *No Cover* condition who believe their matched participant donated to the NAACP (results reported in regression table form in Panel A, Columns 1–3 of Table 3). 27% of respondents in the *No Cover* condition believe their matched participant donated, compared to 35% of respondents in the *Cover* condition ($p = 0.012$). Similarly, the leftmost comparison in Panel B of Figure 3 displays the fraction of participants who deny their matched participant a bonus (results reported in regression table

[32] The full set of experimental instructions is included in Appendix E.2.

form in Panel B, Columns 1–3 of Table 3). 47% of respondents in the *No Cover* condition deny their matched participant a bonus, compared to 40% of respondents in the *Cover* condition ($p = 0.016$). As shown in Table 3, these estimates are stable to the inclusion of controls.

To analyze the open-ended text, we look for the words or phrases of up to three words that are most characteristic of each condition. More precisely, we follow Gentzkow and Shapiro (2010) to calculate Pearson's χ^2 statistic for each phrase.[33] This statistic is higher when the use of the phrase is more asymmetric across treatment conditions and lower for phrases that are used rarely across both conditions. Appendix Table B.11 shows the ten phrases most characteristic of each condition (i.e. with the most positive and the most negative χ^2 scores); consistent with our framework and the treatment effects on the structured measures of inference, we find that respondents in the *Cover* condition are more likely to use phrases related to the article or the associated evidence — for example, "article," "read," "convincing," "increase in crime".[34]

3.3.3 Credibility

To investigate the role of credibility, we run a slightly revised version of Experiment 2 (Auxiliary Experiment 6) with a sample of 506 Democrats and Independents: we instead show respondents the "lower-credibility" versions of the Tweets, as described in Section 3.2.5.[35] We display results in the center comparisons of Panels A and B of Figure 3 and Columns 4–6 of Table 3. While the point estimate of the effect of the rationale on both structured measures of inference remains positive, it is substantially smaller: 30% of respondents in the *No Cover* condition believe their matched partner donated, compared to 33% in the *Cover* condition ($p = 0.58$) and 44% of respondents in the *No Cover* condition deny their matched partner the donation, compared to 42% in the *Cover* condition.[36] While we are underpowered to conclude that the difference in treatment effects between the high-

[33] This statistic is given by: $\chi^2_p = \frac{(n^R_p n^{NR}_{\sim p} - n^{NR}_p n^R_{\sim p})^2}{(n^R_p + n^{NR}_p)(n^R_p + n^R_{\sim p})(n^{NR}_p + n^{NR}_{\sim p})(n^R_{\sim p} + n^{NR}_{\sim p})}$, where n^R_p, n^{NR}_p are the number of times p appears across all responses in the *Cover* condition and *No Cover* condition, respectively, and $n^i_{\sim p}$ is the total number of times a phrase that is *not* p appears in condition i.

[34] These open-ended responses also allow us to mitigate concerns about other potential explanations for our findings: for example, that respondents in the *Cover* condition believed that their matched participant felt pressured by the experimenter to join the campaign and this pressure led them to do so. No respondents mention this or other related confounds.

[35] See Appendix E.15 for experimental instructions.

[36] As shown in Appendix D, our results are unchanged if we reweight responses to match the demographics of the sample in the higher-credibility variation.

credibility and low-credibility wordings is statistically significant, the evidence is consistent with this slightly less credible rationale being substantially less effective.

Our revised experiment also speaks to one of the most common complaints surrounding "political correctness" culture: the alleged tendency of people to "take things out of context". The article prominently lists both Sharkey's academic credentials and, in the first few paragraphs, unequivocally states that "One of the most robust, most uncomfortable findings in criminology is that putting more officers on the street leads to less violent crime." Nonetheless, the revised Tweet appears substantially less effective in shifting inference and reducing social sanctions (suggesting that most respondents do not read the article before deciding whether to sanction their partner). Requirements for dissenters to ensure that no part of their argument can be taken out of context and stripped of accompanying rationales may leave limited scope for expressing nuanced arguments. Conversely, evidence (such as scientific or media articles) may serve as a rationale even if few people actually examine it, so long as it appears compelling at first glance. We discuss implications for the spread of fake and misleading news and for political entrepreneurship in Section 5.

4 Support for Deporting Illegal Immigrants

Our next experiments examine the use and interpretation of rationales among a different population — conservatives — and to justify a different stigmatized position — support for a campaign to immediately deport all illegal Mexican immigrants. We examine our mechanism in this different context for three primary reasons. First, defunding the police is a highly salient but novel policy proposal, and it is thus unclear whether the power of rationales also extends to more "traditional" policy questions, for which there may be more common knowledge about a greater body of evidence and partisan talking points. Second, opposition to defunding the police is likely stigmatized by the in-group (Democrats) but not the out-group (Republicans); in contrast, supporting the immediate deportation of all illegal Mexican immigrants is less stigmatized by the in-group (Republicans), but is highly stigmatized by the out-group (Democrats). This setting thus allows us to examine whether rationales can be used to mitigate social sanctions levied by the out-group as well as from the in-group. Finally, understanding the drivers of anti-immigrant narratives on social media is of direct interest.

As in the previous experiment on the expression of dissent, we study the expression of xenophobia on social media. Given the widespread and growing importance of right-wing media as suppliers of anti-immigrant narratives, we examine a different form of rationale: a thirty-second clip from one of the most popular cable news shows in the US, *Tucker Carlson Tonight*. In the clip, Carlson draws upon statistics from the US Sentencing Commission to argue that illegal immigrants commit violent crimes at substantially higher rates than citizens.[37]

4.1 Experiment 3: Rationales and Pro-Deportation Expression

4.1.1 Sample and experimental design

We conducted our pre-registered Experiment 3 in March 2021 with a sample of Republicans and Independents.[38] We recruited 1,130 participants through Luc.id. After screening out respondents who did not want to join the campaign (as described below), we are left with a final sample of 508 respondents. Our sample is balanced on observables across treatment arms (Appendix Table B.12).

Our experimental design is broadly similar to that of Experiment 1; we provide a diagram in Appendix Figure B.4. As in Experiment 1, respondents log into our survey using their Twitter account and respond to a set of demographic and other background questions. Respondents then view the clip from *Tucker Carlson Tonight*, which is embedded into the survey, and are randomized into the *Cover* condition or the *No Cover* condition. Respondents in the *Cover* condition, but not in the *No Cover* condition, are provided with the URL to the video. We then ask all respondents whether they would like to join a campaign to immediately deport all illegal Mexican immigrants. The survey terminates for respondents who do not join the campaign, leaving us with 517 remaining respondents. Those respondents in the *No Cover* group who do join the campaign are provided the URL to the video. In other words, at this point in the survey, the only difference between conditions is whether respondents are provided with the video URL before (*Cover*) or after (*No Cover*) joining the campaign — though all respondents watch the clip before joining the campaign. As we discuss below, this difference in timing is key to avoiding explicit deception in our experimental manipulation.

The remainder of the experiment is identical in design to Experiment 1, with respondents given the opportunity to schedule the following Tweet in the *Cover* condition:

[37] The clip is available at `https://www.youtube.com/embed/SDdkkTLCUUQ?autoplay=1&amp;controls=0&amp;end=166&amp;fs=0&amp;modestbranding=1&amp;start=113&amp;iv_load_policy=3`.

[38] The full set of experimental instructions is included in Appendix E.3.

> I have joined a campaign to immediately deport all illegal Mexican immigrants. Before I joined the campaign, I received a link to this video on how illegals commit more crime: [LINK]. Sign this petition to immediately deport all illegal Mexicans: [LINK]

Respondents in the *No Cover* condition are presented with an identical Tweet, but with the "Before I joined the campaign..." replaced with "After I joined the campaign...". Although all respondents in fact watched the video before joining the campaign, it is true that respondents in the *Cover* condition *received the link* to the video before joining, while those in the *No Cover* condition received the link after joining.[39] This difference in wording suggests to potential readers of the Tweet that respondents in the *Cover* group had been exposed to the video by Tucker Carlson before joining the campaign — and thus potentially joined because they were convinced by the clip's evidence — while respondents in the *No Cover* group had *not* been exposed before joining the campaign, and thus could not have joined due to the clip. As in Experiment 1, then, this manipulation varies the availability of social cover while fixing the persuasion channel (all respondents are exposed to the same video) and the anticipated persuasion channel (all respondents know their Tweet's readers will be exposed to the video, since it is linked in the Tweet).[40]

4.1.2 Results

The central comparison of Figure 2 displays the results, which we also show in regression table form in Panel A of Table 4. We again find an economically and statistically significant cover effect: 47% of respondents in the *No Cover* condition authorize the Tweet, while 64% of respondents in the *Cover* condition authorize the Tweet ($p < 0.01$, a 0.35 standard deviation effect). The fact that the social cover effect is larger than that estimated in Experiment 1 may reflect that Republicans feel greater stigma in joining a pro-deportation campaign than Democrats feel in joining an anti-defunding campaign (which is also consistent with the lower mean authorization rates in this experiment than in Experiment

[39] One potential concern is that providing a link to respondents in the *Cover* condition, but not in the *No Cover* condition, induces differential selection into the campaign. Because we make the source of the clip obvious, we do not view this as a plausible confound. Indeed, we find no statistically significant difference in selection into the campaign between groups (a 2.6 percentage point difference, $p = 0.474$), and our worst-case estimate under Lee (2009) bounds remains statistically significant at the 1% level.

[40] In principle, we could have used a similar design as Experiment 1: showing the video to respondents both before and after they join the campaign. We concluded that such a manipulation would be less natural for a 30-second video than for a longer article, as in Experiment 1.

1); or that Republicans perceive the *Tucker Carlson* video as a more compelling rationale vis-a-vis their Twitter followers than Democrats perceive the *Washington Post* article vis-a-vis their followers.[41] Turning to treatment effect heterogeneity, we show heterogeneity by demographic characteristics in Column 4 of Appendix Table B.4; we show in Appendix Table B.5 that our estimated treatment effects remain stable when we reweight the sample to match the general population on observables; and we show heterogeneity by partisan affiliation in Panel C of Appendix Table B.7.

4.2 Experiment 4: Interpretation of Pro-Deportation Rationale

Finally, we examine how the availability of the social cover provided by the *Tucker Carlson Tonight* clip shapes an audience's inference about a dissenter's underlying motivations and the resulting social sanctions the dissenter faces.

4.2.1 Sample and experimental design

We conducted our pre-registered Experiment 4 in November 2021 with a sample of 1,082 Democrats and Independents recruited from Prolific.[42] We focus on Democrats and Independents, as anti-immigrant expression is less likely to be stigmatized among Republicans. Our sample is balanced on observables across treatment arms (Appendix Table B.15).

Experiment 4 follows the structure of Experiment 2; Figure B.3 outlines the structure of the experiments (with red text corresponding to Experiment 4). Respondents are informed that they have been matched with a previous survey participant, who joined a campaign to deport all illegal Mexican immigrants. As in Experiment 2, they are then randomized into a *Cover* and a *No Cover* condition: respondents in the *Cover* condition are told that their matched participant authorized the Tweet corresponding to the *Cover* condition of Experiment 3 ("Before I joined the campaign...") whereas respondents in the *No Cover* condition are told that their matched participant authorized the *No Cover* Tweet ("After I joined the campaign..."). Subsequently, they guess whether their matched participant authorized a $5 donation to the US Border Crisis Children's Relief Fund (an organization

[41] In our pre-registered Auxiliary Experiment 8 designed to measure the persuasiveness of the rationale, we find mixed evidence for persuasive effects on private opinions; see Appendix B.2.2 for details and Appendix E.16 for experimental instructions. In a previous working paper (Bursztyn et al., 2020c), we present a series of related pre-registered experiments examining how the availability of an academic rationale affects conservatives' willingness to publicly donate to an anti-immigrant organization. We again find that the rationale increases public anti-immigrant expression.

[42] The full set of experimental instructions is included in Appendix E.4.

that seeks to provide care and basic hygiene items to children along the US–Mexico border) when given the opportunity to do so, and they choose whether or not to deny a \$1 bonus to their matched participant.[43]

4.2.2 Results

The rightmost comparisons of Figure 3 display the fraction of participants in the *Cover* and *No Cover* condition who believe their matched participant donated to the pro-immigrant organization and the corresponding fractions of participants who deny their matched respondent a bonus. 8.5% of respondents in the *No Cover* condition believe their matched participant donated, compared to 13.4% of respondents in the *Cover* condition ($p = 0.01$); 80% of respondents in the *No Cover* condition deny their matched participant a bonus, compared to 74% of respondents in the *Cover* condition ($p = 0.011$). As shown in Panels B and C of Table 4, these estimates are stable to the inclusion of demographic controls.

We plot results from our analysis of open-ended text in Appendix Table B.17, using the same procedure described in Section 3.3.2. As in Experiment 2, respondents in the *Cover* condition are substantially more likely to use words referencing the rationale — "watched a video," "fear mongering," "convinced" — whereas respondents in the *No Cover* condition mention phrases such as "Republican" and "racial". This evidence underscores that rationales shift inference about underlying motives.

5 Discussion and Conclusion

This paper examines how rationales facilitate dissent by lowering the social cost of expressing controversial opinions. In our model, rationales change some people's private views, but they also change an audience's inference about dissenters' motivations and thus can be used to enable dissent. We explore these mechanisms among both liberal and conservative respondents, focusing on a natural setting and outcome: willingness to express dissent on social media. First, we show that liberal respondents are more likely to authorize a Tweet opposing the movement to defund the police when they can credibly ascribe their views to strong scientific evidence. Consistent with our framework, a credible rationale shifts an audience's inference about the respondents and reduces resulting social sanctions. Similarly, conservative respondents are more likely to authorize a Tweet calling for the deportation

[43] We randomized the order of these two different outcomes and detect no significant order effects.

of all illegal immigrants from Mexico — and are seen as less intolerant after doing so — when they can ascribe their views to a Fox News clip.[44]

We now discuss some implications of our framework and empirical results, which may provide fruitful avenues for future research.

Political correctness and the limitations of rationales In a "political correctness" culture, certain rationales cannot be voiced because they are seen as legitimizing dangerous or undesirable causes, and so anyone who uses such a rationale is seen as supporting the cause itself. For example, people who argue for the presence of reverse discrimination against men in labor markets may be seen as sexists: that is, even scientific rationales such as correspondence studies — which may be effective rationales in other settings — may fail to provide a social cover. In some cases, this may be socially desirable: for instance, equating the use of a rationale with sexism may prevent sexist individuals from citing rationales they do not believe or cherry-picking rationales to support their claims. In other cases, political correctness culture may stifle socially important forms of dissenting expression by stigmatizing rationales that would typically be seen as highly credible.[45]

Individuals or institutions seeking to eliminate certain forms of public behavior may use multiple levers to silence dissenters. One lever, explored in Section 3.2.5, is to undermine the credibility of rationales directly. Another lever is to manipulate the real or perceived correlation between knowledge of a rationale and the underlying type, tying the rationale directly to the stigmatized motive.[46] Indeed, in the limit in which only people with stigmatized motives are aware of a certain rationale — e.g., because only they consume the extreme news sources through which the rationale is broadcast — the rationale is

[44] While our experiments explore settings in which there is pressure to express more liberal views — and thus, the rationale supports a more centrist view in Experiment 1 and a more right-wing view in Experiment 3 — our conceptual framework generalizes to any context in which certain types are stigmatized and public expression is informative about type.

[45] The announcement of new ethics requirements in the prominent journal *Nature Human Behavior* highlights this tension (see "Science must respect the dignity and rights of all humans", *Nature Human Behavior*, August 2022): "In some cases. . . potential harms to the populations studied may outweigh the benefit of publication. Academic content that undermines the dignity or rights of specific groups. . . or promotes privileged, exclusionary perspectives raises ethics concerns that may require revisions or supersede the value of publication. . . [but] ensuring that no research is discouraged simply because it may be socially or academically controversial, is as important as preventing harm."

[46] For example, during the Second Red Scare, Joseph McCarthy and his allies explicitly tied several rationales for dissenting with government policy to Communist sympathies. Famously, physicist J. Robert Oppenheimer was stripped of his security clearances when political opponents attributed his opposition to the development of the hydrogen bomb to alleged Soviet loyalties (Cassidy, 2019).

completely ineffective, as to use it is to reveal one's motives with certainty. Tactics to manipulate this real or perceived correlation include disallowing controversial opinions a public platform (e.g., disinviting campus speakers or banning social media accounts), or branding particular media sources or speakers as fringe.[47] Further exploring the conditions under which rationales are most effective, and heterogeneity in the types and sources of rationales which are effective across different groups, is an important direction for future research.[48] For example, evidence that non-credible rationales can backfire, leading to *greater* social sanctions, would have important implications for understanding social dynamics and the supply side of political narratives. Similarly, evidence on how people endogenously acquire rationales and the supply-side implications of such strategic behavior might shed light on both the causes and consequences of increasing polarization in the media.

Political entrepreneurship and populism Populist politicians often scapegoat minorities (Guriev and Papaioannou, 2020; Bursztyn et al., 2022). While the persuasive effects of political propaganda are doubtless important (Adena et al., 2015), propaganda may also generate social cover, enabling supporters to speak their mind more openly and spread the message through their social circle (Satyanath et al., 2017; Caesmann et al., 2021). The strength of this social amplifier depends not only on the number of individuals who hold stigmatized views, but also on the number of individuals who previously *could not express these views.* This may be one reason why the Nazis were able to leverage social networks and associations more effectively than other groups, such as communists: if antisemitism was stigmatized, but relatively common and persistent (Voigtländer and Voth, 2012; Cantoni et al., 2019), then antisemitic Nazi rhetoric generated a large social amplifier. In contrast, blaming economic elites was less stigmatized and thus generated smaller amplifiers.

A more recent rhetorical strategy is dog-whistling: "sending a message to certain potential supporters in such a way as to make it inaudible to others whom it might alienate or deniable for still others who would find any explicit appeal along those lines offensive"

[47] This can also help explain how censorship techniques such as China's "Great Firewall" can be highly effective in repressing discourse unfriendly to the regime, even if citizens can bypass them relatively easily (Chen and Yang, 2019).

[48] Policymakers can also use rationales to affect behavior in non-political settings. For instance, in settings where educational investments are stigmatized (Austen-Smith and Fryer, 2005), providing monetary incentives for exerting educational effort (Levitt et al., 2016) might enable students to attribute educational investments not to academic interest but rather to the incentive. For similar reasons, cold-calling might be preferable to allowing students to volunteer answers.

(Goodin and Saward, 2005). Historians and political scientists have argued that the Republican Party's "Southern Strategy" to win white support in the South was characterized by extensive racial dog-whistling (Haney-López, 2014). In a 1981 interview, Republican strategist and Republican National Committee chairman Lee Atwater described the approach as follows:

> You start out in 1954 by saying, 'N—, n—, n—.' By 1968 you can't say 'n—': that hurts you. Backfires. So you say stuff like forced busing, states' rights and all that stuff. You're getting so abstract now [that] you're talking about cutting taxes, and all these things you're talking about are totally economic things and a byproduct of them is [that] blacks get hurt worse than whites." (Lamis, ed, 1999)

Such dog-whistles generate two types of social cover: one for the politician vis-a-vis the greater public, and one for the politician's supporters vis-a-vis others who disapprove of the statement, allowing them to publicly support the politician and his or her policies without incurring social stigma.

Fake and misleading news on social media Our findings speak to the influence of fake and misleading news on social media. Some recent studies suggest that the persuasive effect of fake and misleading news is limited (Allcott and Gentzkow, 2017; Nyhan, 2018), while others suggest the opposite: that such stories can affect behavior (Barrera et al., 2020) and that individuals may have trouble distinguishing between fake and real news (Angelucci and Prat, 2021) or between facts and opinions (Bursztyn et al., forthcoming). Our results highlight the potential importance of mechanisms beyond persuasion. Specifically, fake and misleading news can generate a social amplifier: rationales that plausibly persuade a small group can change public behavior among a much larger group. This is particularly concerning given that the breadth of rationales, especially those for fringe views, is far greater on social media than on traditional outlets.

Among other platforms, Facebook and Twitter have conducted small-scale experiments evaluating strategies to curtail the spread of misinformation, including warning users before they post an article flagged as fake news and flagging fake or misleading news when it appears on others' timelines. The effects of such interventions are typically modest (Jahanbakhsh et al., 2021). Yet because these changes have not been rolled out at scale, users retain social cover when sharing fake news: they can credibly claim that they did not know the news was fake. Scaling these initiatives to the entire userbase, thus generating common knowledge that all users are warned before posting fake news, would eliminate this cover. For this reason, current (partial equilibrium) estimates of the effects of debunking on users' propensity to share fake news may substantially understate the general equilibrium effects that would be realized if platforms were to fully scale up the feature. At the same time, the evidence from Barrera et al. (2020) emphasizes the importance of platforms' credibility when debunking rationales: when platforms lack credibility, fake and misleading news retains its power to generate social cover for the expression of stigmatized views.

Dynamics Our experiments in this paper investigated a snapshot of the United States between 2021 and 2022. But what are the mechanisms by which social norms surrounding the expression of particular views vary over time and by which particular rationales become more or less effective? Both the credibility of any given rationale and the stigma associated with certain positions (including the rationales and positions we study here) are likely to change over time, either due to strategic manipulation by certain individuals and institutions or due to broader social dynamics. For example, particular topics may become normalized, or particular sources may become delegitimized or associated with stigmatized causes. Rationales that were effective at one time may no longer be effective at a later date — or they may no longer be needed, either because the view has become normalized or because those holding the view care less about the sanctions imposed by the out-groups that would disagree. Understanding these dynamics is an exciting direction for future work.

References

Acemoglu, Daron, Georgy Egorov, and Konstantin Sonin, "A Political Theory of Populism," *Quarterly Journal of Economics*, 2013, *128* (2), 771–805.

Adena, Maja, Ruben Enikolopov, Maria Petrova, Veronica Santarosa, and Ekaterina Zhuravskaya, "Radio and the Rise of The Nazis in Prewar Germany," *Quarterly Journal of Economics*, 2015, *130* (4), 1885–1939.

Ali, S Nageeb and Charles Lin, "Why People Vote: Ethical Motives and Social Incentives," *American Economic Journal: Microeconomics*, 2013, *5* (2), 73–98.

Allcott, Hunt and Matthew A. Gentzkow, "Social Media and Fake News in the 2016 Election," *Journal of Economic Perspectives*, 2017, *31* (2), 211–36.

Angelucci, Charles and Andrea Prat, "Is Journalistic Truth Dead? Measuring How Informed Voters Are about Political News," Working Paper 3593002, Social Science Research Network, 2021.

Austen-Smith, David and Roland G. Fryer, "An Economic Analysis of "Acting White"," *Quarterly Journal of Economics*, 2005, *120* (2), 551–583.

Barrera, Oscar, Sergei Guriev, Emeric Henry, and Ekaterina Zhuravskaya, "Facts, alternative facts, and fact checking in times of post-truth politics," *Journal of Public Economics*, 2020, *182*, 104123.

Bénabou, Roland and Jean Tirole, "Incentives and Prosocial Behavior," *American Economic Review*, 2006, *96* (5), 1652–1678.

_ and _ , "Identity, Morals, and Taboos: Beliefs as Assets," *Quarterly Journal of Economics*, 2011, *126* (2), 805–855.

_ , Armin Falk, and Jean Tirole, "Narratives, Imperatives, and Moral Persuasion," Working Paper 24798, National Bureau of Economic Research, 2020.

Braghieri, Luca, "Political Correctness, Social Image, and Information Transmission," Working paper, 2022.

Bursztyn, Leonardo, Aakaash Rao, Christopher Roth, and David Yanagizawa-Drott, "Opinions as Facts," *Review of Economic Studies*, forthcoming.

_ , Alessandra L. González, and David Yanagizawa-Drott, "Misperceived Social Norms: Women Working Outside the Home in Saudi Arabia," *American Economic Review*, 2020, *110* (10), 2997–3029.

_ and Robert Jensen, "How Does Peer Pressure Affect Educational Investments?," *Quarterly Journal of Economics*, 2015, *130* (3), 1329–1367.

_ , Georgy Egorov, and Stefano Fiorin, "From Extreme to Mainstream: The Erosion of Social Norms," *American Economic Review*, 2020, *110* (11), 3522–48.

_ , _ , Ingar Haaland, Aakaash Rao, and Christopher Roth, "Scapegoating during Crises," *AEA Papers and Proceedings*, 2022, *112*, 151–55.

_ , _ , **Ruben Enikolopov, and Maria Petrova**, "Social media and xenophobia: evidence from Russia," Working Paper 26567, National Bureau of Economic Research, 2019.

_ , **Ingar K Haaland, Aakaash Rao, and Christopher P Roth**, "Disguising prejudice: Popular rationales as excuses for intolerant expression," Technical Report, National Bureau of Economic Research 2020.

Caesmann, Marcel, Bruno Caprettini, Hans-Joachim Voth, David Yanagizawa-Drott et al., "Going Viral: Propaganda, Persuasion and Polarization in 1932 Hamburg," Technical Report, Centre for Economic Policy Research 2021.

Cantoni, Davide, Felix Hagemeister, and Mark Westcott, "Persistence and activation of right-wing political ideology," Working paper, 2019.

Caprettini, Bruno, Marcel Caesmann, Hans-Joachim Voth, and David Yanagizawa-Drott, "Going Viral: Propaganda, Persuasion and Polarization in 1932 Hamburg," Working Paper 16356, Center for Economic and Policy Research, 2021.

Cassidy, David C, *J. Robert Oppenheimer and the American century*, Plunkett Lake Press, 2019.

Chen, Yuyu and David Y. Yang, "The Impact of Media Censorship: 1984 or Brave New World?," *American Economic Review*, 2019, *109* (6), 2294–2332.

Cohn, Nate and Kevin Quealy, "The Democratic Electorate on Twitter Is Not the Actual Democratic Electorate," *The New York Times*, 2019.

Crowne, Douglas P and David Marlowe, "A new scale of social desirability independent of psychopathology.," *Journal of consulting psychology*, 1960, *24* (4), 349.

Cunningham, Tom and Jonathan de Quidt, "Implicit Preferences Inferred from Choice," Working Paper 2709914, Social Science Research Network, 2015.

Dana, Jason, Roberto A. Weber, and Jason Xi Kuang, "Exploiting Moral Wiggle Room: Experiments Demonstrating an Illusory Preference for Fairness," *Economic Theory*, 2007, *33* (1), 67–80.

Dhar, Diva, Tarun Jain, and Seema Jayachandran, "Reshaping adolescents' gender attitudes: Evidence from a school-based experiment in India," *American Economic Review*, 2022, *112* (3), 899–927.

Ekins, Emily E, "Poll: 62% of Americans Say They Have Political Views They're Afraid to Share," Working Paper 3659953, Social Science Research Network, 2020.

Enikolopov, Ruben, Alexey Makarin, and Maria Petrova, "Social media and protest participation: Evidence from Russia," *Econometrica*, 2020, *88* (4), 1479–1514.

_ **and Maria Petrova**, "Chapter 17 - Media Capture: Empirical Evidence," in Simon P. Anderson, Joel Waldfogel, and David Strömberg, eds., *Handbook of Media Economics*, Vol. 1, North-Holland, 2015, pp. 687–700.

Exley, Christine L., "Excusing Selfishness in Charitable Giving: The Role of Risk," *Review of Economic Studies*, 2016, *83* (2), 587–628.

Foerster, Manuel and Joel J van der Weele, "Persuasion, Justification and the Communication of Social Impact," *The Economic Journal*, 2021, *131*, 2887–2919.

Gennaioli, Nicola and Andrei Shleifer, "What Comes to Mind," *Quarterly Journal of Economics*, 2010, *125* (4), 1399–1433.

Gentzkow, Matthew and Jesse M Shapiro, "What drives media slant? Evidence from US daily newspapers," *Econometrica*, 2010, *78* (1), 35–71.

Golman, Russell, "Acceptable Discourse: Social Norms of Beliefs and Opinions," Working paper, 2021.

_ , David Hagmann, and George Loewenstein, "Information Avoidance," *Journal of Economic Literature*, 2017, *55* (1), 96–135.

Goodin, Robert E. and Michael Saward, "Dog Whistles and Democratic Mandates," *The Political Quarterly*, 2005, *76* (4), 471–476.

Grossman, Zachary and Joel J Van Der Weele, "Self-image and willful ignorance in social decisions," *Journal of the European Economic Association*, 2017, *15* (1), 173–217.

Guriev, Sergei and Elias Papaioannou, "The Political Economy of Populism," *Journal of Economic Literature*, 2020.

Haaland, Ingar, Christopher Roth, and Johannes Wohlfart, "Designing Information Provision Experiments," *Journal of Economic Literature*, 2021.

Hamman, John R., George Loewenstein, and Roberto A. Weber, "Self-Interest through Delegation: An Additional Rationale for the Principal-Agent Relationship," *American Economic Review*, 2010, *100* (4), 1826–1846.

Haney-López, Ian, *Dog whistle politics: How coded racial appeals have reinvented racism and wrecked the middle class*, Oxford University Press, 2014.

Jahanbakhsh, Farnaz, Amy X Zhang, Adam J Berinsky, Gordon Pennycook, David G Rand, and David R Karger, "Exploring lightweight interventions at posting time to reduce the sharing of misinformation on social media," *Proceedings of the ACM on Human-Computer Interaction*, 2021, *5* (CSCW1), 1–42.

Kuran, Timur, *Private Truths, Public Lies: The Social Consequences of Preference Falsification*, Cambridge, Massachusetts: Harvard University Press, 1997.

Lacetera, Nicola and Mario Macis, "Social Image Concerns and Prosocial Behavior: Field Evidence from a Nonlinear Incentive Scheme," *Journal of Economic Behavior and Organization*, 2010, *76* (2), 225–237.

Lamis, Alexander P., ed., *Southern Politics in the 1990s*, Baton Rouge: Louisiana State University Press, 1999.

Langer, Ellen J, Arthur Blank, and Benzion Chanowitz, "The mindlessness of ostensibly thoughtful action: The role of "placebic" information in interpersonal interaction," *Journal of Personality and Social Psychology*, 1978, *36* (6), 635.

Lazear, Edward P., Ulrike Malmendier, and Roberto A. Weber, "Sorting in Experiments with Application to Social Preferences," *American Economic Journal: Applied Economics*, 2012, *4* (1), 136–163.

Lee, David S., "Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects," *Review of Economic Studies*, 2009, *76* (3), 1071–1102.

Levitt, Steven D., John A. List, Susanne Neckermann, and Sally Sadoff, "The Behavioralist Goes to School: Leveraging Behavioral Economics to Improve Educational Performance," *American Economic Journal: Economic Policy*, 2016, *8* (4), 183–219.

Levy, Roee and Martin Mattsson, "The Effects of Social Movements: Evidence from# MeToo," Working Paper 3496903, Social Science Research Network, 2021.

Litman, Leib, Jonathan Robinson, and Tzvi Abberbock, "TurkPrime.com: A versatile crowdsourcing data acquisition platform for the behavioral sciences," *Behavior Research Methods*, 2017, *49* (2), 433–442.

Michalopoulos, Stelios and Melanie Meng Xue, "Folklore," *Quarterly Journal of Economics*, 2021, *136* (4), 1993–2046.

Morris, Stephen, "Political correctness," *Journal of Political Economy*, 2001, *109* (2), 231–265.

Müller, Karsten and Carlo Schwarz, "Making America Hate Again? Twitter and Hate Crime Under Trump," Working Paper 3149103, Social Science Research Network, 2018.

Nyhan, Brendan, "Fake News and Bots May Be Worrisome, but Their Political Power Is Overblown," *The New York Times*, 2018.

O'Brien, Sarah, "Employers check your social media before hiring. Many then find reasons not to offer you a job," *CNBC*, 2018.

Ousey, Graham C. and Charis E. Kubrin, "Immigration and Crime: Assessing a Contentious Issue," *Annual Review of Criminology*, 2018, *1* (1), 63–84.

Parker, Kim and Kiley Hurst, "Growing share of Americans say they want more spending on police in their area," *Pew Research Center's Report*, 2021.

Patir, Assaf, Bnaya Dreyfuss, and Moses Shayo, "On the Workings of Tribal Politics," Working Paper 3797290, Social Science Research Network, 2021.

Perez-Truglia, Ricardo and Guillermo Cruces, "Partisan Interactions: Evidence from a Field Experiment in the United States," *Journal of Political Economy*, 2017, *125* (4), 1208–1243.

Reynolds, William M, "Development of reliable and valid short forms of the Marlowe-Crowne Social Desirability Scale," *Journal of clinical psychology*, 1982, *38* (1), 119–125.

Saccardo, Silvia and Marta Serra-Garcia, "Cognitive Flexibility or Moral Commitment? Evidence of Anticipated Belief Distortion," Working paper 3676711, Social Science Research Network, 2020.

Satyanath, Shanker, Nico Voigtländer, and Hans-Joachim Voth, "Bowling for fascism: Social capital and the rise of the Nazi Party," *Journal of Political Economy*, 2017, *125* (2), 478–526.

Shalvi, Shaul, Francesca Gino, Rachel Barkan, and Shahar Ayal, "Self-Serving Justifications: Doing Wrong and Feeling Moral," *Current Directions in Psychological Science*, 2015, *24* (2), 125–130.

Shiller, Robert J, "Narrative economics," *American Economic Review*, 2017, *107* (4), 967–1004.

Thompson, Derek, "Unbundle the Police," *The Atlantic*, 2020.

Voigtländer, Nico and Hans-Joachim Voth, "Persecution perpetuated: the medieval origins of anti-Semitic violence in Nazi Germany," *Quarterly Journal of Economics*, 2012, *127* (3), 1339–1392.

Wood, Thomas and Ethan Porter, "The Elusive Backfire Effect: Mass Attitudes' Steadfast Factual Adherence," *Political Behavior*, 2019, *41* (1), 135–163.

Yanagizawa-Drott, David, "Propaganda and Conflict: Evidence from the Rwandan Genocide," *Quarterly Journal of Economics*, 2014, *129* (4), 1947–1994.

Figures

Figure 1: Experiments 1 and 1R: Experimental design

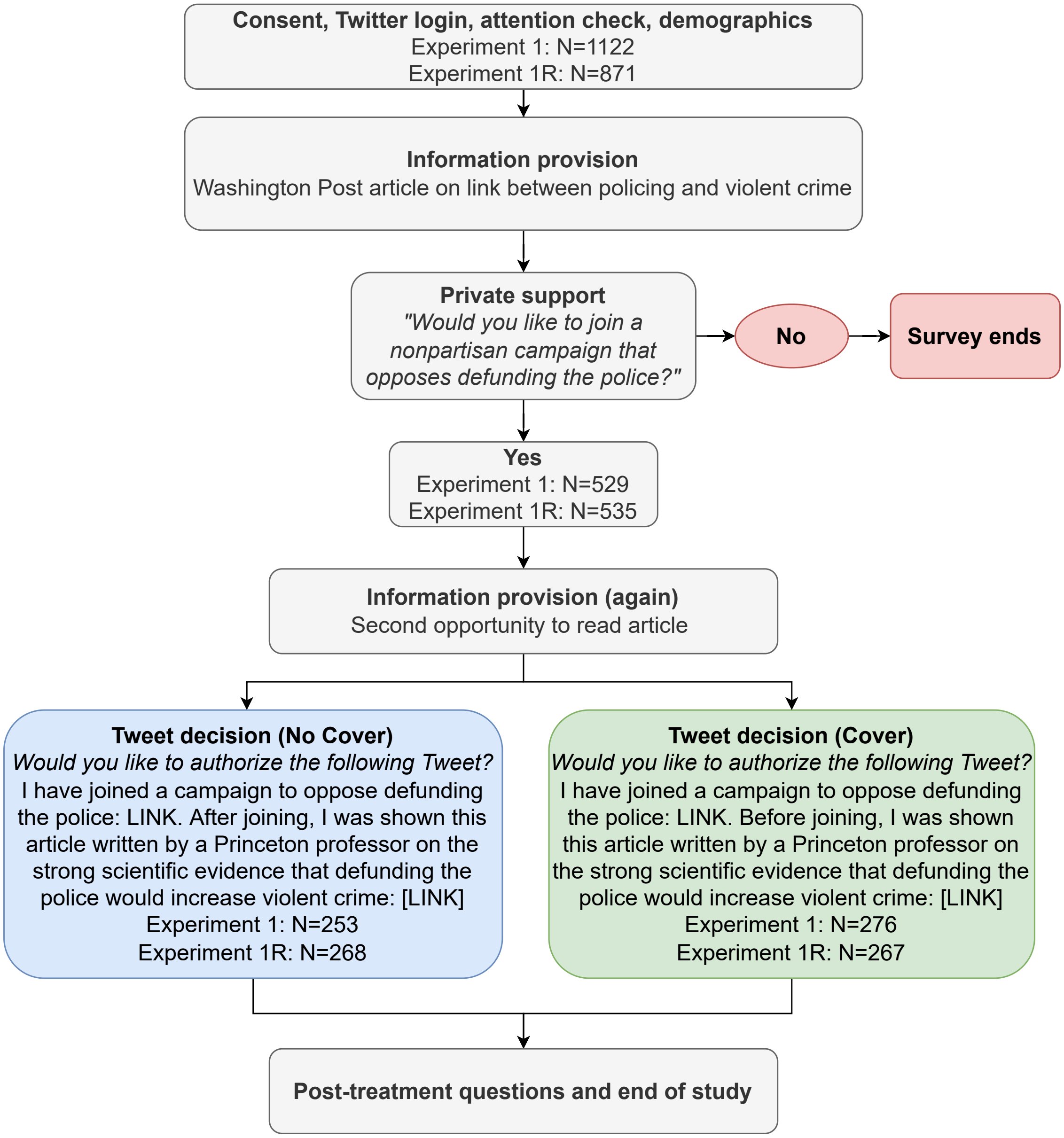
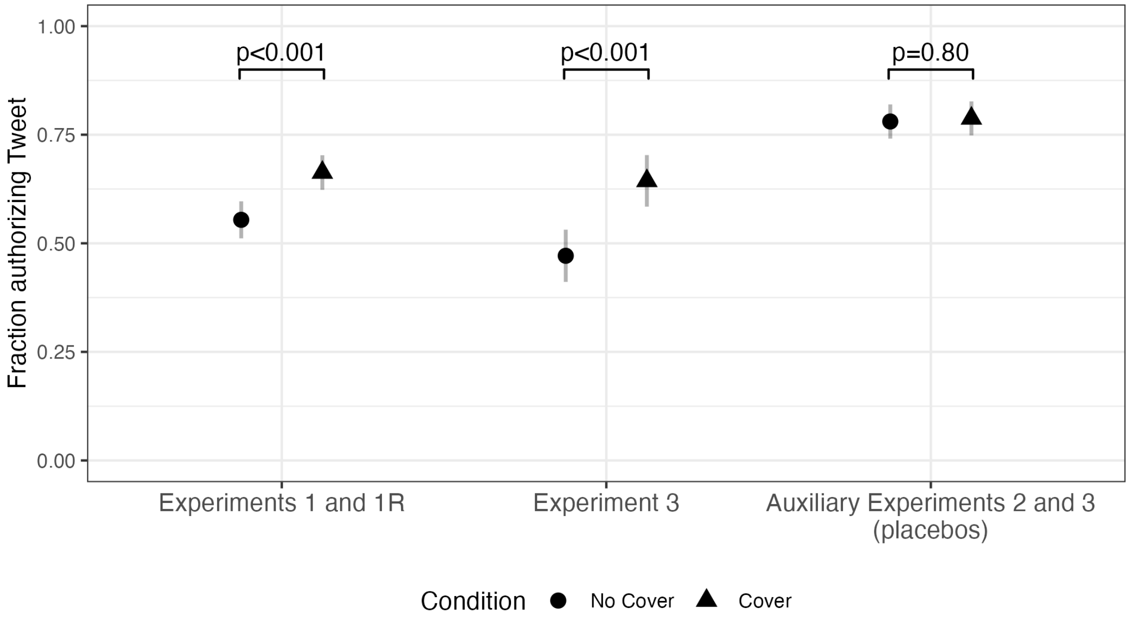
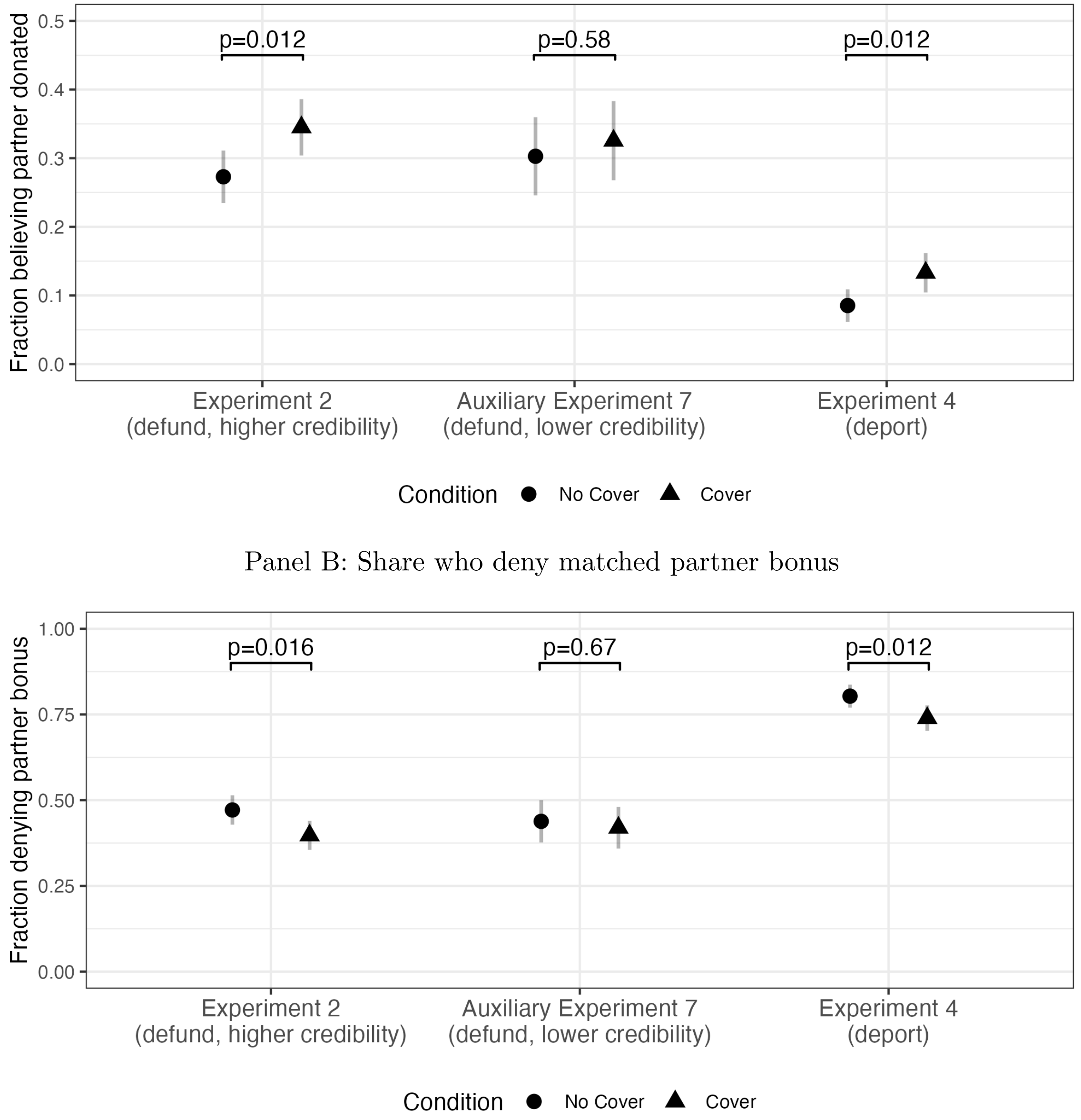

Consent, Twitter login, attention check, demographics
Experiment 1: N=1122
Experiment 1R: N=871
Information provision
Washington Post article on link between policing and violent crime
Private support
"Would you like to join a nonpartisan campaign that opposes defunding the police?"
No
Survey ends
Yes
Experiment 1: N=529
Experiment 1R: N=535
Information provision (again)
Second opportunity to read article
Tweet decision (No Cover)
Would you like to authorize the following Tweet?
I have joined a campaign to oppose defunding the police: LINK. After joining, I was shown this article written by a Princeton professor on the strong scientific evidence that defunding the police would increase violent crime: [LINK]
Experiment 1: N=253
Experiment 1R: N=268
Tweet decision (Cover)
Would you like to authorize the following Tweet?
I have joined a campaign to oppose defunding the police: LINK. Before joining, I was shown this article written by a Princeton professor on the strong scientific evidence that defunding the police would increase violent crime: [LINK]
Experiment 1: N=276
Experiment 1R: N=267
Post-treatment questions and end of study

Figure 2: Fraction authorizing Tweets across experiments



Notes: Figure presents results from Experiment 1 ($n = 523$) and the replication of Experiment 1 ($n = 535$), from Experiment 3 ($n = 508$, and from Auxiliary Experiments 2 ($n = 315$) and 3 ($n = 524$). We plot fraction of respondents authorizing the Tweet, separately by experiment and treatment condition. Error bars indicate 95% confidence intervals. p-values obtained from a two-sample t-test of equality of means.

Figure 3: Interpretation of Tweets



Panel A: Share who believe matched partner donated

Panel B: Share who deny matched partner bonus

Notes: Figure presents results from Experiment 2 ($n = 1040$), Auxiliary Experiment 7 ($n = 506$), and Experiment 4 ($n = 1082$). Panel A plots the fraction of respondents who believe their matched partner donated to the organization in question: the NAACP for Experiment 2 and Auxiliary Experiment 7, and the USBCCRF for Experiment 4. Panel B plots the fraction of respondents who deny their partner a $1 bonus. Error bars indicate 95% confidence intervals. p-values obtained from a two-sample t-test of equality of means.

Tables

Table 1: Willingness to post anti-defunding Tweet

	Scheduled Tweet					
	Main	Replication	Pooled	Main	Replication	Pooled
	(1)	(2)	(3)	(4)	(5)	(6)
Cover	0.124***	0.092**	0.108***	0.119***	0.096**	0.108***
	(0.042)	(0.042)	(0.030)	(0.042)	(0.042)	(0.030)
No Cover mean	0.568	0.541	0.554	0.568	0.541	0.554
Controls	No	No	No	Yes	Yes	Yes
Observations	523	535	1,058	523	535	1,058
R^2	0.017	0.009	0.014	0.062	0.056	0.037

Notes: Table reports results from Experiment 1 and the replication of Experiment 1 (Experiment 1R). The dependent variable is an indicator taking value 1 if the respondent chose to schedule the post. Columns 1 and 4 limit to the sample from Experiment 1; Columns 2 and 5 limit to the sample from Experiment 1R; Columns 3 and 6 pool the two samples and include experiment fixed effects. Controls include age, age squared, a set of race indicators, a Hispanic indicator, a male indicator, and a set of education indicators. Robust standard errors are reported. *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively.

Table 2: Interpreting effects of rationale on willingness to post anti-defunding Tweet

	Mean		Treatment effect	
	No Cover	Cover	Coef. (s.e.)	p-value
Panel A: Replication of Exp. 1 (Exp. 1R, $n = 535$)				
Respondent believes Tweet is...				
Misleading	0.04	0.01	-0.02 (0.01)	0.11
Misleading about timing	0.00	0.00	0.00 (0.00)	—
Panel B: Rainforest placebo (Aux. Exp. 2, $n = 315$)				
Scheduled post	0.83	0.79	-0.04 (0.04)	0.38
Panel C: Daylight saving placebo (Aux. Exp. 3, $n = 524$)				
Scheduled post	0.75	0.79	0.04 (0.04)	0.34
Respondent believes Tweet is...				
Misleading	0.06	0.07	0.02 (0.02)	0.48
Misleading about timing	0.00	0.00	0.00 (0.00)	—
Panel D: Anticipated persuasion (Aux. Exp. 4, $n = 501$)				
Estimated share persuaded	25.34	27.23	1.90 (2.12)	0.37
Panel E: Open-ended motive elicitation (Aux. Exp. 5, $n = 402$)				
E.1: Primary motives: respondent mentions...				
Social cover	0.15	0.25	0.10 (0.04)	0.02
Anticipated persuasion	0.07	0.06	-0.01 (0.02)	0.67
Information	0.57	0.50	-0.07 (0.05)	0.13
E.2: Potential confounds: respondent mentions...				
Unnatural	0.01	0.01	0.01 (0.01)	0.32
Misleading	0.00	0.00	0.00 (0.00)	—
Signaling	0.00	0.00	0.00 (0.00)	—
Experimenter demand	0.00	0.00	0.00 (0.00)	—
Panel F: Credibility manipulation (Aux. Exp. 6, $n = 1,017$)				
F.1: Hypothetical willingness to post				
Willing to post (high cred.)	0.57	0.67	0.11 (0.04)	0.02
Willing to post (low cred.)	0.57	0.62	0.05 (0.04)	0.21
F.2: Beliefs about social sanctions				
Share denying bonus (high cred.)	53.14	48.05	-5.09 (2.31)	0.03
Share denying bonus (low cred.)	53.99	53.00	-0.99 (2.06)	0.63

Notes: In Panels A and C, "Misleading" and "Misleading about timing" are indicators for whether the respondent found the Tweet misleading and whether the respondent found the Tweet misleading specifically about when they read the article relative to joining the campaign. In Panel D, DV is the respondent's guess about the percentage of their followers who would join the campaign if they saw the Tweet. In Panel E, DVs are indicators for whether the respondent's motive falls in each of the categories. p-values obtained from a two-sided t-test of equality of means.

Table 3: Inference about and social sanctions toward matched anti-defunding respondent

	Higher-credibility		Lower-credibility	
	(1)	(2)	(3)	(4)
Panel A:	*Belief partner donated*			
Cover	0.072**	0.072**	0.023	0.023
	(0.029)	(0.029)	(0.041)	(0.042)
No Cover mean	0.273	0.273	0.303	0.303
R^2	0.006	0.024	0.001	0.034
Panel B:	*Denied bonus to partner*			
Cover	−0.074**	−0.074**	−0.019	−0.028
	(0.031)	(0.031)	(0.044)	(0.044)
No Cover mean	0.471	0.471	0.438	0.438
R^2	0.006	0.040	0.0004	0.059
Controls	No	Yes	No	Yes
Observations	1,040	1,037	506	506

Notes: Table reports results from Experiment 2 (columns 1–3) and Auxiliary Experiment 6 (columns 4–6). The dependent variable in Panel A is an indicator taking value 1 if the respondent reports believing that his or her matched partner donated to the US Border Crisis Children's Relief Fund. The dependent variable in Panel B is an indicator taking value 1 if the respondent denied his or her matched partner a $1 bonus. Controls include age, age squared, a set of race indicators, a Hispanic indicator, a male indicator, and a set of education indicators. Robust standard errors are reported. *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively.

Table 4: Expression and interpretation of pro-deportation Tweet

	Experiment 3	
Panel A:	*Scheduled Tweet*	
Cover	0.172***	0.179***
	(0.044)	(0.044)
No Cover mean	0.471	0.471
Observations	508	508
R^2	0.030	0.071
	Experiment 4	
Panel B:	*Belief partner donated*	
Cover	0.048**	0.049**
	(0.019)	(0.019)
No Cover mean	0.085	0.085
Observations	1,080	1,079
R^2	0.006	0.033
Panel C:	*Denied bonus to partner*	
Cover	−0.064**	−0.064**
	(0.026)	(0.026)
No Cover mean	0.803	0.803
Observations	1,080	1,079
R^2	0.006	0.024
Controls	No	Yes

Notes: Panel A presents the results of Experiment 3, in which the dependent variable is an indicator taking value 1 if the respondent chose to schedule the post. Panels B and C present the results of Experiment 4. The dependent variable in Panel B is an indicator taking value 1 if the respondent reports believing that his or her matched partner donated to the US Border Crisis Children's Relief Fund (USBCCRF). The dependent variable in Panel C is an indicator taking value 1 if the respondent denied his or her matched partner a $1 bonus. Controls include age, age squared, a set of race indicators, a Hispanic indicator, a male indicator, and a set of education indicators. Robust standard errors are reported. *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively.

Online Appendix: Not for publication

Our supplementary material is structured as follows. Appendix A provides proofs of all theoretical results in Section 2 and discusses the case of polarizing rationales. Appendix B.1 provides supporting material for the experiments presented in Section 3. Appendix B.2 provides supporting material for the experiments presented in Section 4. Appendix C discusses the ethical considerations underlying all experimental designs. Finally, Appendix E provides the instruments for all experiments described in the paper.

A Theoretical Results

A.1 Proof of Proposition 1

We first prove that for random variable t distributed with c.d.f. $H(\cdot)$ and p.d.f. $h(\cdot)$,

$$\frac{d}{d\tau}\mathbb{E}\left(t \mid t > \tau\right) \leq 1.$$

Let $z_\tau = t - \tau$ be a family of random variables indexed by τ; we need to show that

$$\mathbb{E}\left(z_\tau \mid z_\tau \geq 0\right)$$

is non-increasing in τ. Denoting the c.d.f. of z_τ by $F_\tau(\cdot)$ and its p.d.f. by $f_\tau(\cdot)$, we have

$$\mathbb{E}\left(z_\tau \mid z_\tau \geq 0\right) = \frac{1}{1 - F_\tau(0)} \int_0^{+\infty} y f_\tau(y)\, dy.$$

The integral may be rewritten as

$$\begin{aligned} \int_0^{+\infty} y f_\tau(y)\, dy &= \int_0^{+\infty} f_\tau(y) \left(\int_0^y 1 dx\right) dy = \int_0^{+\infty} \int_0^y f_\tau(y)\, dx dy \\ &= \int_0^{+\infty} \int_x^{+\infty} f_\tau(y)\, dy dx = \int_0^{+\infty} \left(1 - F_\tau(x)\right) dx, \end{aligned}$$

where we used Fubini's theorem to change the order of integration.

Note that $F_\tau(x) = \Pr\left(z_\tau \leq x\right) = \Pr\left(t \leq x + \tau\right) = H(x + \tau)$. We therefore have

$$\mathbb{E}\left(z_\tau \mid z_\tau \geq 0\right) = \int_0^{+\infty} \frac{1 - F_\tau(x)}{1 - F_\tau(0)} dx = \int_0^{+\infty} \frac{1 - H(x + \tau)}{1 - H(\tau)} dx.$$

The integrand is non-increasing in τ pointwisely (i.e., for any fixed $x \geq 0$), because

$$\begin{aligned} \frac{d}{d\tau}\left(\frac{1-H(x+\tau)}{1-H(\tau)}\right) &= \frac{h(\tau)(1-H(x+\tau))-h(x+\tau)(1-H(\tau))}{(1-H(\tau))^2} \\ &= \frac{1-H(x+\tau)}{1-H(\tau)}\left(\frac{h(\tau)}{1-H(\tau)}-\frac{h(x+\tau)}{1-H(x+\tau)}\right) \leq 0, \quad \text{(A1)} \end{aligned}$$

because the first term is positive and the second is nonpositive due to monotone hazard rate property. This proves that $\mathbb{E}(z_\tau \mid z_\tau \geq 0)$ is non-increasing in τ, and thus $\frac{d}{d\tau}\mathbb{E}(t \mid t > \tau) \leq 1$.

Now, for any fixed social cost S, type t_i would choose $d_i = 1$ if $t_i > \frac{1}{\beta}S - w_0$ and would choose $d_i = 0$ if the opposite inequality holds. Thus, every equilibrium is characterized by a threshold τ. This threshold τ satisfies the condition

$$G(\tau) = -w_0, \tag{A2}$$

where

$$G(\tau) = \tau - \frac{\gamma}{\beta}\left(\mathbb{E}(t_i \mid t_i > \tau) - \bar{t}\right). \tag{A3}$$

Since, as we proved, $\frac{d}{d\tau}\mathbb{E}(t_i \mid t_i > \tau) \leq 1$ and $\gamma < \beta$, the $G(\tau)$ is strictly increasing in τ, and furthermore

$$\frac{d}{d\tau}G(\tau) \geq 1 - \frac{\gamma}{\beta} > 0.$$

This shows that equation (A2) has a unique solution τ_0, and the corresponding social cost equals $S_0 = \gamma\left(\mathbb{E}(t_i \mid t_i > \tau_0) - \bar{t}\right)$. This completes the proof. ■

A.2 Proof of Proposition 2

Since the distributions are normal, the posterior of citizen i is given by the usual formula

$$w_1 = \mathbb{E}(w \mid s) = w_0\frac{\sigma_\varepsilon^2}{\sigma_w^2+\sigma_\varepsilon^2} + s\frac{\sigma_w^2}{\sigma_w^2+\sigma_\varepsilon^2}.$$

We have

$$w_1 - w_0 = \frac{\sigma_w^2}{\sigma_w^2+\sigma_\varepsilon^2}(s-w_0),$$

so $w_1 > w_0$.

In the private signal case, the social cost is fixed at S_0, and therefore $S_{priv} = S_0$. Given that, the threshold τ_{priv} satisfies

$$\tau_{priv} - \frac{1}{\beta}S_0 = -w_1.$$

Therefore, $\tau_{priv} = \frac{1}{\beta} S_0 - w_1 = \tau_0 - G(\tau_0) - w_1 = \tau_0 + (w_0 - w_1)$, so $\tau_{priv} < \tau_0$.

Now consider the public signal case. In this case, the equilibrium takes the form of a threshold τ_{pub} that satisfies

$$G(\tau_{pub}) = -w_1, \tag{A4}$$

where $G(\cdot)$ is defined in (A3). Since $\frac{d}{d\tau} G(\tau) > 0$ and $-w_1 < -w_0$, we have $\tau_{pub} < \tau_0$. Furthermore, since $\frac{d}{d\tau} G(\tau) < 1$, the difference $\tau_0 - \tau_{pub} > w_1 - w_0 > 0$, and since we showed above that $w_1 - w_0 = \tau_0 - \tau_{priv}$, we have $\tau_{priv} > \tau_{pub}$. As for the social cost, it now equal $S_{pub} = \gamma(\mathbb{E}(t_i \mid t_i > \tau_{pub}) - \bar{t}) < \gamma(\mathbb{E}(t_i \mid t_i > \tau_0) - \bar{t})$, so $S_{pub} < S_0 = S_{priv}$.

Now consider an increase in σ_ε^2, which implies a decrease in w_1. We have $\frac{d}{dw_1}(\tau_0 - \tau_{priv}) = \frac{d}{dw_1}(w_1 - w_0) = 1$, and a decrease in w_1 brings the difference $\tau_0 - \tau_{priv}$ closer to 0. We also have $\frac{d}{d\tau} G(\tau) < 1$, and since τ_{pub} solves (A4), by the implicit function theorem we have $\frac{d}{dw_1}(\tau_{pub}) < -1$. Consequently, $\frac{d}{dw_1}(\tau_0 - \tau_{pub}) > 1$ and $\frac{d}{dw_1}(\tau_{priv} - \tau_{pub}) > 0$, which implies that these differences also decrease as w_1 decreases. As for the social cost, S_{priv} does not depend on w_1 and S_{pub} is increasing in τ_{pub} and thus decreasing in w_1, so $S_{priv} - S_{pub}$ is increasing in w_1. A increase in σ_ε^2, therefore, decreases $S_{priv} - S_{pub}$.

Lastly, as $\gamma \to 0$, the all social costs vanish, and then $\tau_{priv} \to -w_1$ and $\tau_{pub} \to -w_1$, and w_1 does not depend on γ. This shows that the difference between τ_{priv} and τ_{pub} vanishes as well, which completes the proof. ■

A.3 Polarizing Rationales

To formally show that a rationale does not have to be persuasive to be effective, consider the following extension. Consider a polarizing rationale that different people interpret differently. Specifically, suppose that share μ of citizens get a high signal $s_h > w_0$ (with the corresponding posterior $w_h > w_0$) and share $1 - \mu$ get a low signal $s_l < w_0$ (and their posterior is $w_l < w_0$).

Proposition 3. *Suppose that*

$$\mu\left(H(\tau_0) - H(\tau_0 - (w_h - w_0))\right) \geq (1 - \mu)\left(H(\tau_0 + (w_0 - w_l)) - H(\tau_0)\right), \tag{A5}$$

where τ_0 is the equilibrium threshold in the basic model (Proposition 1). Then the ex ante probability that citizen i chooses $d_i = 1$ is higher than in the basic model, and the equilibrium social cost is lower.

In other words, if the mass of people who are persuaded to choose $d_i = 1$ by high signal s_h (holding the social cost fixed) is at least as large as the mass of people who are dissuaded from doing so by low signal s_l, then the social cost of choosing $d_i = 1$ goes down in equilibrium, and more people do so in equilibrium. Intuitively, the audience now faces the inference problem: citizen i may have chosen $d_i = 1$ either because t_i is high, or because he got a high signal s_h. More precisely, the set of citizens who would choose to support change C now contains some types with $t_i < \tau_0$ (moderates who got a high signal s_h) and lacks some types with $t_i > \tau_0$ (extremists who got a low signal s_l). As long as the share of the former is not too small, the posterior of t_i conditional on choosing $d_i = 1$ goes down. As a result, more citizens choose $d_i = 1$ and face a lower social cost for doing so. This result is not knife-edge: it applies even if somewhat more people are dissuaded.

A.3.1 Proof

Notice first that our assumption of rational expectation of t_i conditional on $d_i = 1$ allows us to bypass the discussion of whether members of the audience get signals s_l, s_h, or both. Rational expectation can be formed in practice if people had prior interactions with those who choose $d_i = 1$ and learned their type, which allows them to form a correct expectation in equilibrium about individuals who choose $d_i = 1$ with a given piece of evidence. An alternative way is to assume that the audience is sophisticated and understands the whole signal structure, but does not know which signal citizen i got, and faces the signal decomposition problem as a result.

In what follows, we let $\bar{t} = 0$ to save on notation, but the proof goes through for any $\bar{t}$. We start by establishing the uniqueness of equilibrium in this case. Let $\bar{S}$ be the social cost of choosing $d_i = 1$ in a hypothetical equilibrium. Then the citizen would choose $d_i = 1$ if $t_i > \frac{1}{\beta}\bar{S} - w_h$ following signal s_h and if $t_i > \frac{1}{\beta}\bar{S} - w_l$ following signal s_l. This implies that there are two thresholds, τ_h and τ_l, that satisfy $\tau_l - \tau_h = w_h - w_l$. Denote $\bar{\tau} = \frac{1}{\beta}\bar{S} - w_0$; then $\tau_h = \bar{\tau} + w_0 - w_h$ and $\tau_l = \bar{\tau} + w_0 - w_l$. From now on we describe the equilibrium in terms of $\bar{\tau}$.

We will use the following probabilities. We denote

$$p\left(x, y\right) = \mu\left(1 - H\left(x\right)\right) + \left(1 - \mu\right)\left(1 - H\left(y\right)\right),$$

so

$$p\left(\bar{\tau} + w_0 - w_h, \bar{\tau} + w_0 - w_l\right) = p\left(\frac{1}{\beta}\bar{S} - w_h, \frac{1}{\beta}\bar{S} - w_l\right)$$

is the probability of choosing $d_i = 1$ if the citizen faces social cost $\bar{S}$. We also let

$$q\left(x, y\right) = \frac{\mu\left(1 - H\left(x\right)\right)}{p\left(x, y\right)},$$

so $q\left(\bar{\tau} + w_0 - w_h, \bar{\tau} + w_0 - w_l\right)$ is the equilibrium conditional probability that citizen i got signal s_h conditional on choosing $d_i = 1$.

Define the function

$$\begin{aligned}\bar{S}\left(z\right) &= \gamma q\left(z + w_0 - w_h, z + w_0 - w_l\right)\mathbb{E}\left(t_i \mid t_i > z + w_0 - w_h\right) \\ &\quad + \gamma\left(1 - q\left(z + w_0 - w_h, z + w_0 - w_l\right)\right)\mathbb{E}\left(t_i \mid t_i > z + w_0 - w_l\right).\end{aligned}$$

In equilibrium characterized by $\bar{\tau}$, the social cost of choosing $d_i = 1$ equals $\bar{S}(\bar{\tau})$. Given the above, thresholds $\tau_h = \bar{\tau} + w_0 - w_h$ and $\tau_l = \bar{\tau} + w_0 - w_l$ are equilibrium thresholds for choosing $d_i = 1$ after getting signals s_h and s_l, respectively, if and only if $\bar{\tau}$ solves the equation

$$\bar{\tau} - \frac{1}{\beta}\bar{S}(\bar{\tau}) = -w_0. \tag{A6}$$

Let us show that $\frac{d}{dz}\frac{1}{\gamma}\bar{S}(z) \leq 1$. Indeed, from the proof of Proposition 1, we have

$$\begin{aligned} \frac{d}{dz}\mathbb{E}(t_i \mid t_i > z + w_0 - w_h) &\leq 1; \\ \frac{d}{dz}\mathbb{E}(t_i \mid t_i > z + w_0 - w_l) &\leq 1. \end{aligned}$$

Furthermore,

$$\mathbb{E}(t_i \mid t_i > z + w_0 - w_l) > \mathbb{E}(t_i \mid t_i > z + w_0 - w_h).$$

Lastly, we have

$$\begin{aligned} q(z + w_0 - w_h, z + w_0 - w_l) &= \frac{\mu(1 - H(z + w_0 - w_h))}{\mu(1 - H(z + w_0 - w_h)) + (1-\mu)(1 - H(z + w_0 - w_l))} \\ &= \frac{1}{1 + \frac{1-\mu}{\mu}\frac{1-H(z+w_0-w_l)}{1-H(z+w_0-w_h)}}. \end{aligned}$$

Now,

$$\frac{d}{dz}\frac{1 - H(z + w_0 - w_l)}{1 - H(z + w_0 - w_h)} = \frac{d}{du}\frac{1 - H(u + (w_h - w_l))}{1 - H(u)} \leq 0,$$

where we denoted $u = z + w_0 - w_h$ and used the calculation (A1) from the proof of Proposition 1. This immediately implies that $\frac{d}{dz}q(z + w_0 - w_h, z + w_0 - w_l) \geq 0$. Summing up, we have

$$\begin{aligned} \frac{d}{dz}\frac{1}{\gamma}\bar{S}(z) &= q(z + w_0 - w_h, z + w_0 - w_l)\frac{d}{dz}\mathbb{E}(t_i \mid t_i > z + w_0 - w_h) \\ &\quad + (1 - q(z + w_0 - w_h, z + w_0 - w_l))\frac{d}{dz}\mathbb{E}(t_i \mid t_i > z + w_0 - w_l) \\ &\quad + \left(\frac{d}{dz}q(z + w_0 - w_h, z + w_0 - w_l)\right) \\ &\quad \times (\mathbb{E}(t_i \mid t_i > z + w_0 - w_h) - \mathbb{E}(t_i \mid t_i > z + w_0 - w_l)). \end{aligned}$$

Notice that the sum of the first two lines does not exceed 1 (since both derivatives do not exceed 1), and term on the third line is positive and the one on the fourth is negative, so their product is negative. This proves that $\frac{d}{dz}\frac{1}{\gamma}\bar{S}(z) \leq 1$. Now, as in the proof of Proposition 1 this implies that the equation (A6) has a unique solution $\bar{\tau}$, which proves the uniqueness of equilibrium in this case.

Let us now show that in this solution, $\bar{\tau} < \tau_0$ and $\bar{S}(\bar{\tau}) < S_0$, where S_0 is the equilibrium social cost in the basic model. To do this, it is sufficient to show that $\bar{S}(\tau_0) < S_0$. Indeed, this would imply that

$$\tau_0 - \frac{1}{\beta}\bar{S}(\tau_0) > \tau_0 - \frac{1}{\beta}S_0 = -w_0,$$

and since $\bar{\tau}$ satisfies (A6) and the function $x - \frac{1}{\beta}\bar{S}(x)$ is increasing, we would get $\bar{\tau} < \tau_0$. Then we would get

$$\bar{S}(\bar{\tau}) = \beta(\bar{\tau} + w_0) < \beta(\tau_0 + w_0) = S_0,$$

as required. So, to complete the proof, we need to show that $\bar{S}(\tau) < S_0$.

In the light of condition (A5) and by continuity of $H(\cdot)$, there exists $\hat{w}_h \in (0, w_h)$ such that

$$\mu(H(\tau_0) - H(\tau_0 - (\hat{w}_h - w_0))) = (1-\mu)(H(\tau_0 + (w_0 - w_l)) - H(\tau_0)).$$

Let $\hat{S}$ denote the value

$$\begin{aligned}\hat{S} &= \gamma q(\tau_0 + w_0 - \hat{w}_h, \tau_0 + w_0 - w_l)\mathbb{E}(t_i \mid t_i > \tau_0 + w_0 - \hat{w}_h) \\ &\quad + \gamma(1 - q(\tau_0 + w_0 - \hat{w}_h, \tau + w_0 - w_l))\mathbb{E}(t_i \mid t_i > \tau_0 + w_0 - w_l);\end{aligned}$$

in other words, the expression for $\hat{S}$ is analogous to $\bar{S}(\tau)$, except that w_h is replaced by $\hat{w}_h$.

We now show that $\bar{S}(\tau) < \hat{S} < S_0$. To prove the first inequality, we use some algebra to establish that

$$\frac{1}{\gamma}\bar{S}(\tau) = (1-\rho)\frac{1}{\gamma}\hat{S} + \rho\mathbb{E}(t_i \mid t_i \in (\tau_0 + w_0 - w_h, \tau_0 + w_0 - \hat{w}_h)),$$

where

$$\rho = q(\tau_0 + w_0 - w_h, \tau_0 + w_0 - w_l)\frac{H(\tau_0 + w_0 - \hat{w}_h) - H(\tau_0 + w_0 - w_h)}{1 - H(\tau_0 + w_0 - w_h)}.$$

Since $\rho > 0$ and $\frac{1}{\gamma}\hat{S} < \mathbb{E}(t_i \mid t_i \in (\tau_0 + w_0 - w_h, \tau_0 + w_0 - \hat{w}_h))$ as the former is an expectation taken over values to the right of $\tau_0 + w_0 - \hat{w}_h$ while the latter expectation is taken over values to the left of that point, we get $\bar{S}(\tau) < \hat{S}$.

Let us now prove that $\hat{S} < S_0$. Spelling out $q(\tau_0 + w_0 - \hat{w}_h, \tau_0 + w_0 - w_l)$ and expectations in the definition of $\hat{S}$, we have

$$\begin{aligned}\frac{1}{\gamma}\left(S_0-\hat{S}\right) &= \frac{\int_{\tau_0}^{\infty} xh\left(x\right)dx}{1-H\left(\tau_0\right)}\\ &\quad -\frac{\mu\int_{\tau_0+w_0-\hat{w}_h}^{\infty} xh\left(x\right)dx+\left(1-\mu\right)\int_{\tau_0+w_0-w_l}^{\infty} xh\left(x\right)dx}{\mu\left(1-H\left(\tau_0+w_0-\hat{w}_h\right)\right)+\left(1-\mu\right)\left(1-H\left(\tau_0+w_0-w_l\right)\right)}.\end{aligned}$$

Notice that by the definition of $\hat{w}_h$ the denominators in both terms are equal, hence $S_0-\hat{S}$ has the same sign as

$$\begin{aligned}&\int_{\tau_0}^{\infty} xh\left(x\right)dx-\left(\mu\int_{\tau_0+w_0-\hat{w}_h}^{\infty} xh\left(x\right)dx+\left(1-\mu\right)\int_{\tau_0+w_0-w_l}^{\infty} xh\left(x\right)dx\right)\\ =\;&\left(1-\mu\right)\int_{\tau_0}^{\tau_0+w_0-w_l} xh\left(x\right)dx-\mu\int_{\tau_0+w_0-\hat{w}_h}^{\tau_0} xh\left(x\right)dx\\ =\;&\left(1-\mu\right)\left(H\left(\tau_0+w_0-w_l\right)-H\left(\tau_0\right)\right)\mathbb{E}\left(t_i\mid t_i\in\left(\tau_0,\tau_0+w_0-w_l\right)\right)\\ &-\mu\left(H\left(\tau_0\right)-H\left(\tau_0+w_0-\hat{w}_h\right)\right)\mathbb{E}\left(t_i\mid t_i\in\left(\tau_0+w_0-\hat{w}_h,\tau_0\right)\right).\end{aligned}$$

Since the coefficients in front of the expectations in the last two lines are the same (again, by the choice of $\hat{w}_h$), the sign of this expression is the same as the sign of

$$\mathbb{E}\left(t_i\mid t_i\in\left(\tau_0,\tau_0+w_0-w_l\right)\right)-\mathbb{E}\left(t_i\mid t_i\in\left(\tau_0+w_0-\hat{w}_h,\tau_0\right)\right),$$

which is positive, because the first term is greater than τ_0 and the second is less than that. Therefore, $\hat{S}<S_0$.

We have thus proved that $\bar{S}\left(\tau\right)<\hat{S}<S_0$ which, as we showed earlier, implies the results stated. This completes the proof. ■

B Additional Details on Experiments

Table B.1: Overview of data collections

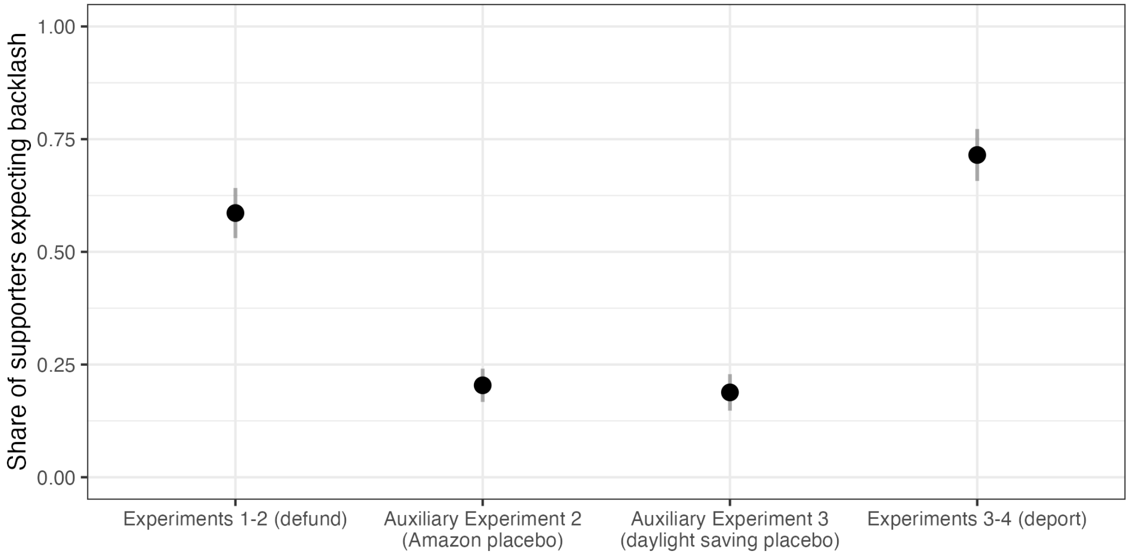
		Sample	Size	PR ID	Source	Date
Panel A: Main experiments						
Exp. 1	Willingness to post anti-defunding Tweet	D, I	523	0008432	L, C	10/21
Exp. 1R	Replication of Experiment 1	D	535	0010269	L, C	10/22
Exp. 2	Interpretation of anti-defunding Tweet	D, I	1,040	0005462	P	11/21
Exp. 3	Willingness to post pro-deportation Tweet	R, I	508	0007379	L	3/21
Exp. 4	Interpretation of pro-deportation Tweet	D, I	1,082	0005462	P	11/21
Panel B: Auxiliary collections						
Aux. Survey	Anticipated social sanctions	D, I, R	505	111262	L	11/22
Aux. Exp. 1	Persuasiveness of anti-defunding article	D, I	1,008	0008624	P	12/21
Aux. Exp. 2	Placebo: willingness to post pro-conservation Tweet	D, I, R	315	—	L, C	12/21
Aux. Exp. 3	Placebo: willingness to post anti-DST Tweet	D, R	524	0005479	L, C	11/22
Aux. Exp. 4	Anticipated persuasiveness of anti-defunding Tweet	D, I	501	—	P	11/21
Aux. Exp. 5	Motives underlying anti-defunding posting decision	D, I	402	—	P	1/22
Aux. Exp. 6	Effect of credibility on anti-defunding posting decision	D, I	1,017	—	L	7/22
Aux. Exp. 7	Interpretation of low-credibility anti-defunding Tweet	D, I	506	—	P	11/21
Aux. Exp. 8	Persuasiveness of pro-deportation video	R, I	2,012	0008624	L, P	12/21

Notes: "Sample" column indicates whether sample consisted of Democrats (D), Independents (I), and/or Republicans (R). "PR ID" column lists the pre-registration IDs for pre-registered collections (AEA RCT registry for all experiments; AsPredicted for the survey); experiments with no corresponding ID were not pre-registered. "Source" column indicates whether respondents were recruited through Luc.id (L), CloudResearch (C), and/or Prolific (P).

B.1 Anti-Defunding Experiments

B.1.1 Auxiliary Survey

Figure B.1: Anticipated social sanctions



Notes: For each cause, Appendix Figure B.1 plots the share of respondents who anticipate they would face "Significant" or "Strong" social backlash if they were to support the cause on Twitter. The sample for each cause is limited to respondents who privately support the cause. Error bars indicate 95% confidence intervals.

B.1.2 Experiment 1: Additional Figures and Tables

Table B.2: Experiment 1: Balance of covariates

	Overall		Cover	No Cover	*p*-value
	mean	std. dev.	mean	mean	(C = NC)
Panel A:			Experiment 1		
Age	39.81	15.13	39.25	40.42	0.38
Black	0.21	0.41	0.23	0.20	0.33
Asian	0.08	0.27	0.07	0.08	0.77
White	0.67	0.47	0.67	0.68	0.82
Hispanic	0.19	0.39	0.16	0.21	0.14
Male	0.59	0.49	0.58	0.59	0.76
High school diploma	0.98	0.16	0.97	0.98	0.90
Bachelors degree	0.44	0.50	0.42	0.45	0.48
Independent	0.26	0.44	0.23	0.30	0.07
Ideology	-0.56	0.98	-0.55	-0.57	0.86
Panel B:			Replication of Experiment 1		
Age	35.33	11.84	35.41	35.25	0.87
Black	0.21	0.41	0.21	0.20	0.73
Asian	0.07	0.25	0.06	0.07	0.74
White	0.71	0.45	0.71	0.72	0.90
Hispanic	0.07	0.25	0.06	0.08	0.31
Male	0.53	0.50	0.54	0.52	0.57
High school diploma	0.99	0.09	0.99	0.99	1.00
Bachelors degree	0.56	0.50	0.57	0.54	0.46
Ideology	-1.07	0.79	-1.02	-1.13	0.13

Notes: *p*-values based on robust standard errors reported. Ideology is coded from -2 (very liberal) to 2 (very conservative).

Table B.3: Experiment 1: Sample representativeness

	Defund	Pew (Inds and Dems)
Panel A:	Experiment 1	
Age	39.81	45.86
Black	0.21	0.18
White	0.67	0.67
Asian	0.08	0.05
Hispanic	0.19	0.15
Male	0.59	0.46
High school diploma	0.98	0.89
Bachelors degree or higher	0.44	0.35
	Defund	Pew (Dems)
Panel B:	Replication of Experiment 1	
Age	35.33	46.67
Black	0.21	0.26
White	0.71	0.58
Asian	0.07	0.05
Hispanic	0.07	0.17
Male	0.53	0.39
High school diploma	0.99	0.86
Bachelors degree or higher	0.56	0.36

Notes: Table displays mean characteristics, comparing the experimental sample with the 2018 Pew Research Center's American Trends Panel, Wave 39. Attriters are dropped from sample.

Table B.4: Heterogeneity by demographic characteristics (Tweetability experiments)

	Treatment effect: coef. (s.e.)			
	Defund	Rainforest	DST	Deport
Panel A: Heterogeneity by age				
Under 25	0.13 (0.08)	-0.25 (0.16)	-0.33 (0.15)	0.35 (0.35)
25-40	0.08 (0.04)	0.06 (0.06)	0.08 (0.06)	0.14 (0.08)
40-60	0.12 (0.05)	-0.13 (0.08)	0.10 (0.06)	0.28 (0.06)
60+	0.13 (0.10)	0.08 (0.12)	-0.05 (0.08)	-0.02 (0.09)
Panel B: Heterogeneity by race and ethnicity				
White, not Hispanic	0.10 (0.04)	-0.00 (0.05)	0.05 (0.04)	0.17 (0.05)
Non-white or Hispanic	0.13 (0.05)	-0.11 (0.08)	0.01 (0.07)	0.24 (0.14)
Panel C: Heterogeneity by gender				
Male	0.12 (0.04)	0.03 (0.07)	0.05 (0.05)	0.21 (0.06)
Female	0.10 (0.05)	-0.10 (0.06)	0.02 (0.06)	0.14 (0.06)
Panel D: Heterogeneity by education				
Bachelor's degree or higher	0.08 (0.04)	-0.02 (0.08)	0.03 (0.06)	0.01 (0.07)
No bachelor's degree or higher	0.13 (0.04)	-0.06 (0.05)	0.04 (0.05)	0.27 (0.05)

Notes: Table reports treatment effects for Experiments 1 and 1R (pooled, with experiment fixed effects), Auxiliary Experiment 2 (rainforest placebo), Auxiliary Experiment 8 (daylight saving placebo), and Experiment 3. Panels A–D split the sample by age, race and ethnicity, gender, and education, respectively. Robust standard errors are reported.

Table B.5: Willingness to post anti-defunding or pro-deportation Tweet (reweighted estimates)

	Scheduled Tweet			
	Experiment 1		Experiment 3	
	(1)	(2)	(3)	(4)
Cover	0.111*** (0.030)	0.119*** (0.030)	0.135*** (0.044)	0.142*** (0.044)
No Cover mean	0.554	0.554	0.471	0.471
Controls	No	Yes	No	Yes
Observations	1,058	1,058	508	508
R^2	0.013	0.042	0.018	0.065

Notes: Columns 1–2 report results from Experiments 1 and 1R. Columns 3–4 report results from Experiment 3. The dependent variable is an indicator taking value 1 if the respondent chose to schedule the post. Table reweights the sample to match the population on age, race, Hispanic identity, gender, and education. Controls include age, age squared, a set of race indicators, a Hispanic indicator, a male indicator, and a set of education indicators. Robust standard errors are reported. *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively.

Table B.6: Experiment 1R: heterogeneity by Twitter characteristics

	Scheduled Tweet		
Panel A: Number of followers			
	0–25 (1)	26–100 (2)	100+ (3)
Cover	0.102 (0.102)	0.067 (0.067)	0.109 (0.073)
No Cover mean	0.529	0.577	0.495
Observations	106	207	188
R^2	0.153	0.118	0.139
Panel B: Percentage of audience opposed			
	0–30% (1)	30–70% (2)	70–100% (3)
Cover	−0.073 (0.091)	0.238*** (0.060)	0.019 (0.095)
No Cover mean	0.662	0.493	0.471
Observations	121	258	121
R^2	0.109	0.136	0.154
Controls	Yes	Yes	Yes

Notes: Table reports heterogeneity in estimated treatment effects in Experiment 1R. The dependent variable is an indicator taking value 1 if the respondent chose to schedule the post. Panel A splits the sample based on the self-reported number of followers. Panel B splits the sample based on the self-reported share of audience who would oppose defunding. Respondents whose self-reported number of followers are inconsistent with actual number of followers excluded. Controls include age, age squared, a set of race indicators, a Hispanic indicator, a male indicator, and a set of education indicators. Robust standard errors are reported. *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively.

Table B.7: Heterogeneity by partisan affiliation (Experiments 1–4)

	All (1)	Partisan (2)	Independent (3)
Panel A: Experiments 1 and 1R (pooled)			
		Scheduled Tweet	
Cover	0.109***	0.129***	0.016
	(0.030)	(0.032)	(0.091)
No Cover mean	0.554	0.538	0.649
Panel B: Experiment 2			
		Belief partner donated	
Cover	0.072**	0.047	0.153**
	(0.029)	(0.032)	(0.061)
No Cover mean	0.273	0.265	0.298
		Denied bonus to partner	
Cover	−0.074**	−0.055	−0.118*
	(0.031)	(0.035)	(0.061)
No Cover mean	0.471	0.482	0.435
Panel C: Experiment 3			
		Scheduled Tweet	
Cover	0.179***	0.237***	−0.081
	(0.044)	(0.049)	(0.104)
No Cover mean	0.471	0.457	0.516
Panel D: Experiment 4			
		Belief partner donated	
Cover	0.049**	0.048**	0.045
	(0.019)	(0.021)	(0.041)
No Cover mean	0.0853	0.0744	0.118
		Denied bonus to partner	
Cover	−0.064**	−0.068**	−0.076
	(0.026)	(0.030)	(0.053)
No Cover mean	0.803	0.811	0.779

Notes: Column 2 restricts the sample to Democrats (Panels A, B, and D) or Republicans (Panel C). Panel A includes experiment fixed effects. Controls include age, age squared, a set of race indicators, a Hispanic indicator, a male indicator, and a set of education indicators. Robust standard errors are reported. *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively.

B.1.3 Auxiliary Experiment 1: Persuasiveness of Defunding Rationale

We conducted this pre-registered experiment in December 2021 with a sample of 1,008 Democrats and Independents recruited from Prolific. After completing a set of demographic questions, respondents assigned to the treatment group read Sharkey's article in the *Washington Post*, while respondents assigned to the control group did not read the article. They then respond to the following two questions: "How do you think increasing funding for the police would affect violent crime?" and "Do you think that funding for the police should be increased, decreased, or stay the same?". We code both questions from -2 ("Decreased a lot" and "Strongly decrease violent crime", respectively) to 2 ("Increased a lot" and "Strongly decrease violent crime", respectively).

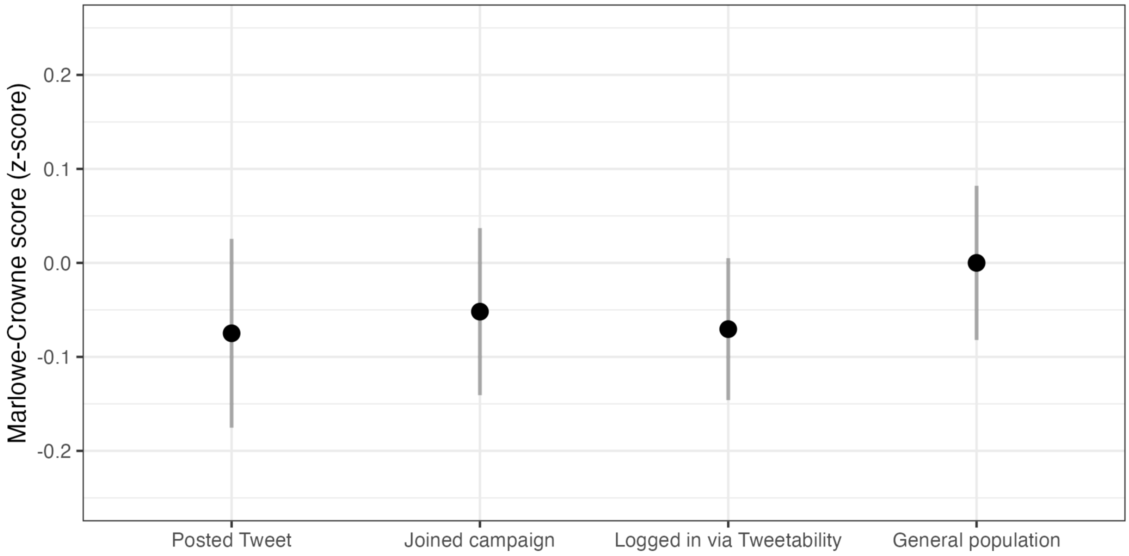
Table B.8 displays results, with Columns 1–3 corresponding to the first measure and Columns 4–6 corresponding to the second measure. We find a significant effect on both measures, though effects are weaker for policy preferences and are no longer significant once we control for demographics and partisan affiliation.

Table B.8: Persuasive effects of anti-defunding article

	Belief		*Policy preference*	
	(1)	(2)	(3)	(4)
Provided article	-0.243^{***}	-0.250^{***}	0.129^{*}	0.117^{*}
	(0.056)	(0.055)	(0.071)	(0.068)
No Article mean	0.036	0.036	-0.638	-0.638
Controls	No	Yes	No	Yes
Observations	1,004	1,003	1,004	1,003
R^2	0.018	0.085	0.003	0.107

Notes: Table reports results from Auxiliary Experiment 1. The dependent variable in Columns 1–3 is the respondent's reported belief as to the effect of increasing funding for the police on violence crime, coded between -2 ("Strongly decrease violent crime") and 2 ("Strongly increase violent crime"). The dependent variable in Columns 4–6 is the respondent's reported preference for changing police funding, ranging from -2 ("Decreased a lot") to 2 ("Increased a lot"). Controls include age, age squared, a set of race indicators, a Hispanic indicator, a male indicator, and a set of education indicators. Robust standard errors are reported. *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively.

Figure B.2: Marlowe-Crowne scores by sample



Notes: Appendix Figure B.2 plots the mean Marlowe-Crowne score (a measure of respondents' concern for social approval, with higher values indicating greater concern for social approval) among each of four different samples: a general population sample, the sample which logged in via Tweetability in Auxiliary Experiment 3, the sample which joined the campaign in Auxiliary Experiment 3, and the sample which posted the Tweet in Auxiliary Experiment 3. Scores have been scaled to mean zero and standard deviation one. Error bars indicate 95% confidence intervals.

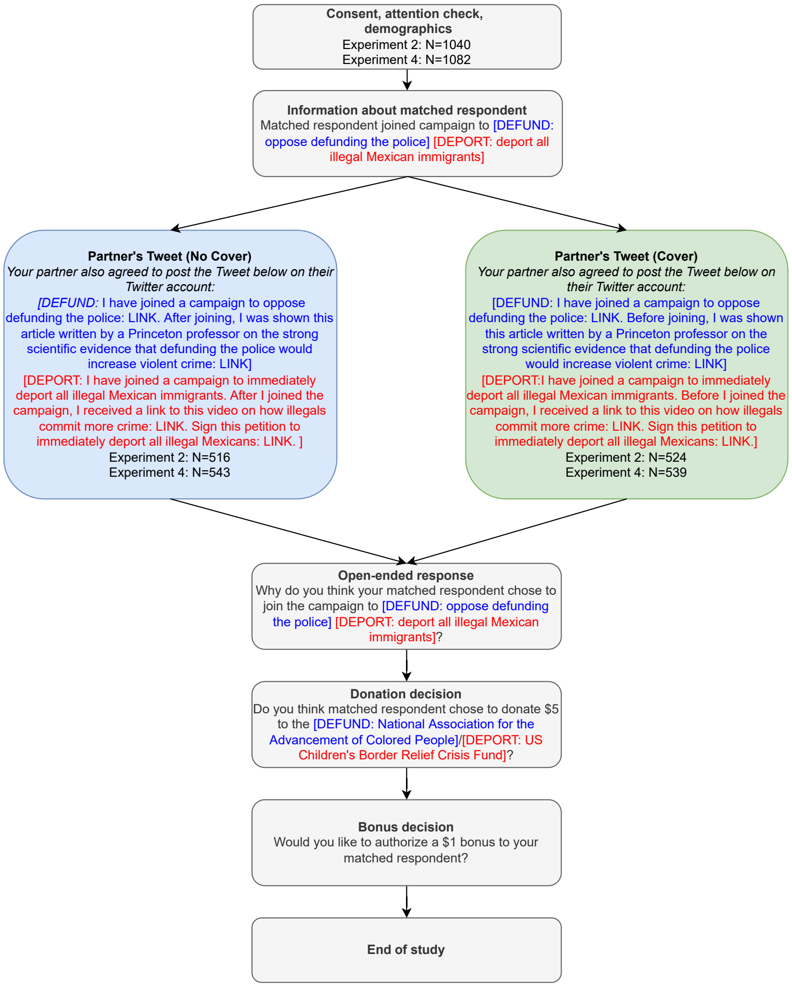
B.1.4 Auxiliary Experiment 4: Anticipated Persuasion Experiment

We conducted this experiment in November 2021 with a sample of 501 Democrats and Independents recruited from Prolific. Only Democrats and Independents with Twitter accounts were eligible to take the survey. After completing a set of demographic questions, respondents read Sharkey's article in the *Washington Post*. As in Experiment 1, respondents are asked if they would like to join the campaign to oppose the movement to defund the police, only those who indicate that they would like to join the campaign proceed with the experiment, and those who do proceed are given a chance to re-read the article. They are then randomly shown either the *Cover* or the *No Cover* Tweet from Experiment 1 and are asked: "Suppose you posted the Tweet above on your account. If you had to guess, what percentage of people who saw your Tweet would choose to join the campaign to oppose defunding the police?"

Panel B of Table 2 displays results. Reassuringly, we find no significant difference between the anticipated persuasiveness of the Tweets, suggesting that differential posting rates are instead driven by changes in anticipated stigma.

B.1.5 Experiment 2: Additional Figures and Tables

Figure B.3: Experiments 2 and 4: flow of inference design


Consent, attention check, demographics
Experiment 2: N=1040
Experiment 4: N=1082
Information about matched respondent
Matched respondent joined campaign to [DEFUND: oppose defunding the police] [DEPORT: deport all illegal Mexican immigrants]
Partner's Tweet (No Cover)
Your partner also agreed to post the Tweet below on their Twitter account:
[DEFUND: I have joined a campaign to oppose defunding the police: LINK. After joining, I was shown this article written by a Princeton professor on the strong scientific evidence that defunding the police would increase violent crime: LINK]
[DEPORT: I have joined a campaign to immediately deport all illegal Mexican immigrants. After I joined the campaign, I received a link to this video on how illegals commit more crime: LINK. Sign this petition to immediately deport all illegal Mexicans: LINK.]
Experiment 2: N=516
Experiment 4: N=543
Partner's Tweet (Cover)
Your partner also agreed to post the Tweet below on their Twitter account:
[DEFUND: I have joined a campaign to oppose defunding the police: LINK. Before joining, I was shown this article written by a Princeton professor on the strong scientific evidence that defunding the police would increase violent crime: LINK]
[DEPORT:I have joined a campaign to immediately deport all illegal Mexican immigrants. Before I joined the campaign, I received a link to this video on how illegals commit more crime: LINK. Sign this petition to immediately deport all illegal Mexicans: LINK.]
Experiment 2: N=524
Experiment 4: N=539
Open-ended response
Why do you think your matched respondent chose to join the campaign to [DEFUND: oppose defunding the police] [DEPORT: deport all illegal Mexican immigrants]?
Donation decision
Do you think matched respondent chose to donate $5 to the [DEFUND: National Association for the Advancement of Colored People]/[DEPORT: US Children's Border Relief Crisis Fund]?
Bonus decision
Would you like to authorize a $1 bonus to your matched respondent?
End of study

Notes: Experiments 2 and 4 have identical structures, so we present both experiments jointly. Blue text corresponds to Experiment 2, studying opposition to the movement to defund the police; red text corresponds to Experiment 4, studying support for immediately deporting all illegal Mexican immigrants.

Table B.9: Experiment 2: Balance of covariates

	Overall		Cover	No Cover	*p*-value
	mean	std. dev.	mean	mean	(C = NC)
Age	30.73	11.26	30.69	30.76	0.91
Black	0.07	0.26	0.09	0.06	0.06
Asian	0.08	0.28	0.09	0.08	0.59
White	0.77	0.42	0.77	0.78	0.56
Hispanic	0.11	0.31	0.09	0.13	0.06
Male	0.37	0.48	0.38	0.36	0.52
High school diploma	1.00	0.05	1.00	1.00	0.55
Bachelors degree	0.57	0.49	0.56	0.58	0.52
Independent	0.25	0.43	0.25	0.24	0.52
Ideology	-1.13	0.77	-1.09	-1.17	0.07

Notes: *p*-values based on robust standard errors reported. Ideology is coded from -2 (very liberal) to 2 (very conservative).

Table B.10: Experiment 2: Sample representativeness

	Defund	Pew (Inds and Dems)
Age	30.73	45.86
Black	0.07	0.18
White	0.77	0.67
Asian	0.08	0.05
Hispanic	0.11	0.15
Male	0.37	0.46
High school diploma	1.00	0.89
Bachelors degree or higher	0.57	0.35

Notes: Table displays mean characteristics, comparing the experimental sample with the 2018 Pew Research Center's American Trends Panel, Wave 39. Attriters are dropped from sample.

Table B.11: Experiment 2: most characteristic words of each condition

Most characteristic of *Cover*	**Most characteristic of *No Cover***
article	things
read	family
written	away
line	possibly
convincing	general
idea	money
increase in crime	understand
article written	programs
washington	interactions
washington post	family members

Notes: Table reports the ten 1-3 word phrases from Experiment 2 with the most positive χ^2 values and the most negative χ^2 values (that is, most characteristic of the *Cover* and *No Cover* conditions, respectively).

B.2 Anti-Immigrant Experiments

B.2.1 Experiment 3: Additional Figures and Tables

Figure B.4: Experiment 3: design

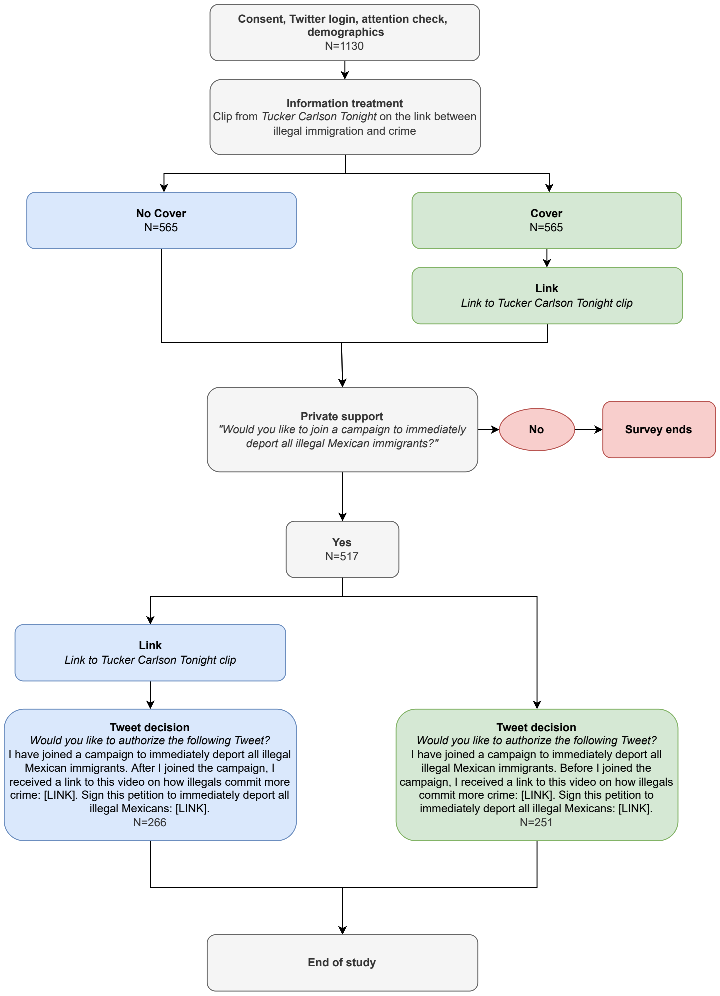

Consent, Twitter login, attention check, demographics
N=1130
Information treatment
Clip from Tucker Carlson Tonight on the link between illegal immigration and crime
No Cover
N=565
Cover
N=565
Link
Link to Tucker Carlson Tonight clip
Private support
"Would you like to join a campaign to immediately deport all illegal Mexican immigrants?"
No
Survey ends
Yes
N=517
Link
Link to Tucker Carlson Tonight clip
Tweet decision
Would you like to authorize the following Tweet?
I have joined a campaign to immediately deport all illegal Mexican immigrants. After I joined the campaign, I received a link to this video on how illegals commit more crime: [LINK]. Sign this petition to immediately deport all illegal Mexicans: [LINK].
N=266
Tweet decision
Would you like to authorize the following Tweet?
I have joined a campaign to immediately deport all illegal Mexican immigrants. Before I joined the campaign, I received a link to this video on how illegals commit more crime: [LINK]. Sign this petition to immediately deport all illegal Mexicans: [LINK].
N=251
End of study

Table B.12: Experiment 3: Balance of covariates

	Overall		Cover	No Cover	*p*-value
	mean	std. dev.	mean	mean	(C = NC)
Age	49.23	13.55	48.51	49.90	0.25
Black	0.01	0.11	0.01	0.01	0.95
Asian	0.02	0.12	0.02	0.02	0.94
White	0.95	0.21	0.95	0.96	0.73
Hispanic	0.06	0.25	0.05	0.08	0.27
Male	0.50	0.50	0.49	0.52	0.54
High school diploma	0.99	0.08	1.00	0.99	0.60
Bachelors degree	0.38	0.49	0.34	0.42	0.07
Independent	0.22	0.41	0.20	0.24	0.29
Ideology	0.97	0.92	0.97	0.97	0.94

Notes: *p*-values based on robust standard errors reported. Ideology is coded from -2 (very liberal) to 2 (very conservative).

Table B.13: Experiment 3: Sample representativeness

	Deport	Pew (Inds and Reps)
Age	49.23	47.17
Black	0.01	0.05
White	0.95	0.83
Asian	0.02	0.03
Hispanic	0.06	0.11
Male	0.50	0.52
High school diploma	0.99	0.93
Bachelors degree or higher	0.38	0.31

Notes: Table displays mean characteristics, comparing the experimental sample with the 2018 Pew Research Center's American Trends Panel, Wave 39. Attriters are dropped from sample.

B.2.2 Auxiliary Experiment 8: Persuasiveness of Deportation Rationale

We conducted a first pre-registered experiment in December 2021 with a sample of 1,008 Republicans recruited from Prolific. After completing a set of demographic questions, respondents assigned to the treatment group viewed the clip from *Tucker Carlson Tonight*, while respondents assigned to the control group did not view the clip. They then indicated their agreement with the following two statements: "Illegal immigrants are not much more likely to commit serious crimes than U.S. citizens" (beliefs) and "The US should immediately deport all illegal Mexican immigrants" (policy preference). We code both questions from -2 ("Strongly disagree") to 2 ("Strongly agree").

Panel A of Table B.14 displays results. While we found significant effects on the beliefs outcome, we found no treatment effects on the policy preference outcome. Two logistical problems complicate interpretation of this result. First, when setting up the survey, we forgot to exclude respondents from some previous experiments which included the video. Thus, some respondents in the *Control* condition had seen the video in previous experiments. Second, there was a highly limited sample of Republicans available on Prolific (fewer than 2000 who met our screening criteria), and we had to pay a higher than usual rate in order to meet our pre-registered sample size. This potentially induced selection into the survey.

We thus ran the same experiment on Luc.id, with the same sample restrictions. , with Columns 1–3 corresponding to the first measure and Columns 4–6 corresponding to the second measure. We find a significant effect on both measures, with an effect size of around 0.12 standard deviations for the first outcome and 0.18 standard deviations for the second outcome.

Overall, we take the evidence for the effects of the clip on persuasion as mixed.

Table B.14: Persuasive effects of *Tucker Carlson Tonight* video

	Belief		*Policy preference*	
	(1)	(2)	(3)	(4)
Panel A:	*Prolific Sample*			
Provided article	0.536***	0.541***	0.001	−0.019
	(0.063)	(0.063)	(0.074)	(0.073)
No Article mean	0.301	0.301	0.541	0.541
Observations	1,004	1,004	1,004	1,004
R^2	0.067	0.091	0.00000	0.062
Panel B:	*Lucid Sample*			
Provided article	0.751***	0.743***	0.177**	0.179**
	(0.066)	(0.066)	(0.074)	(0.073)
No Article mean	0.251	0.251	0.652	0.652
Observations	1,004	1,002	1,004	1,002
R^2	0.115	0.141	0.006	0.061
Controls	No	Yes	No	Yes

Notes: Table reports results from Auxiliary Experiment 7. The dependent variable in Columns 1–3 is the respondent's reported agreement with the statement "Illegal immigrants are more likely to commit serious crimes than US citizens," coded between -2 ("Strongly disagree") and 2 ("Strongly agree"). The dependent variable in Columns 4–6 is the respondent's reported agreement with the statement "The US should immediately deport all illegal Mexican immigrants," ranging from -2 ("Strongly disagree") to 2 ("Strongly agree"). Panel A uses the sample from Prolific, and Panel B uses the sample from Luc.id. Controls include age, age squared, a set of race indicators, a Hispanic indicator, a male indicator, and a set of education indicators. Robust standard errors are reported. *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively.

B.2.3 Experiment 4: Additional Figures and Tables

Table B.15: Experiment 4: Balance of covariates

	Overall		Cover	No Cover	p-value
	mean	std. dev.	mean	mean	(C = NC)
Age	31.60	11.91	32.16	31.05	0.13
Black	0.07	0.25	0.06	0.08	0.40
Asian	0.10	0.30	0.09	0.11	0.26
White	0.77	0.42	0.79	0.75	0.16
Hispanic	0.12	0.32	0.11	0.13	0.34
Male	0.48	0.50	0.49	0.47	0.43
High school diploma	1.00	0.07	0.99	1.00	0.66
Bachelors degree	0.59	0.49	0.59	0.59	0.88
Independent	0.25	0.44	0.26	0.25	0.92
Ideology	-1.12	0.79	-1.08	-1.16	0.09

Notes: p-values based on robust standard errors reported. Ideology is coded from -2 (very liberal) to 2 (very conservative).

Table B.16: Experiment 4: Sample representativeness

	Deport	Pew (Inds and Dems)
Age	31.60	45.86
Black	0.07	0.18
White	0.77	0.67
Asian	0.10	0.05
Hispanic	0.12	0.15
Male	0.48	0.46
High school diploma	1.00	0.89
Bachelors degree or higher	0.59	0.35

Notes: Table displays mean characteristics, comparing the experimental sample with the 2018 Pew Research Center's American Trends Panel, Wave 39. Attriters are dropped from sample.

Table B.17: Experiment 4: most characteristic words of each condition

Most characteristic of *Cover*	**Most characteristic of *No Cover***
watched	people
video	politics
link	lot
fear	racial
media	usa
influenced	illegal immigrant
watched a video	day
decision	liberal
fear mongering	respondent
convinced	republican

Notes: Table reports the ten 1-3 word phrases from Experiment 4 with the most positive χ^2 values and the most negative χ^2 values (that is, most characteristic of the *Cover* and *No Cover* conditions, respectively).

C Ethical Considerations

Understanding dissenting expression is of great social importance. Identifying the drivers of xenophobic expression is crucial in designing policies best-suited to curbing it, while understanding barriers to dissenting expression in situations where such expression is desirable — for example, speaking out against unjust practices or systems — may help design contexts with lower such barriers.

Nonetheless, ethically conducting revealed-preference experiments on dissenting expression — particularly xenophobic expression — requires balancing three often contradictory objectives: avoiding explicitly deceiving respondents, avoiding compromising respondents' privacy, and avoiding increasing public xenophobic expression. In this section, we provide a more detailed explanation of how our experimental designs balance these objectives. All experiments obtained approval from multiple Institutional Review Boards.

C.1 Considerations related to information provision (Experiments 3–4)

The raw numbers pertaining to violent crime cited in the *Tucker Carlson Tonight* clip that we provide to respondents in Experiments 3–4 are taken from the U.S. Sentencing Commission and are factually correct. Nonetheless, the clip paints an incomplete picture of the academic literature, which generally finds null or negative effects of illegal immigration on violent crime. Although we do not endorse this evidence, we nonetheless debrief all respondents at the end of the study, providing them with an accessible academic overview of the link between illegal immigration and violent crime (Ousey and Kubrin, 2018) and a list of further readings.

C.2 Considerations related to privacy and deception (Experiments 1, 1R, and 3 and Auxiliary Experiments 2–3)

Given that our mechanism examines the effect of perceived social stigma on behavior, it is crucial that respondents in Experiments 1 and 3 believe that their decisions will be visible to others. Although our experiments avoid explicit deception, protecting participants' privacy and avoiding starting a political campaign in these contexts required us to mislead respondents. We distinguish between the ethical and practical problems associated with deception (the latter relating to concerns about subject pool contamination), addressing the first concern in this section and the second in Section C.3.

Twitter login All respondents were required to log in via their Twitter accounts to the "Tweetability" app we created. This app is governed by the Twitter API's terms of service and has the second most restrictive set of permissions among the three application scopes Twitter provides ("Read" and "Write"). That is, the app does not have access to users' passwords, messages, or account settings, but it is able to post Tweets from the users' accounts. We do not use this functionality in any way, and no information that could compromise users' accounts is ever accessed or downloaded. We explicitly inform respondents of the app's permissions in transparent language and give them the option to end the survey if they are uncomfortable granting the app these permissions. We also inform respondents that the app's data, including the tokens that give us access to post on their accounts, will be deleted by no later than August 1, 2021 (Experiment 3), December 1, 2021 (Experiment 1), March 1, 2022 (Auxiliary Experiment 2), and December 1, 2022 (Experiment 1R and Auxiliary Experiment 3). Tokens were indeed deleted immediately after collection.

Twitter posts Our key outcome is whether respondents are willing to post a Tweet including a link to a petition to immediately deport all illegal Mexican immigrants. We were not willing to consider designs that asked respondents to actually post such Tweets. We thus asked them to "schedule" their Tweet for the future (using the Tweetability app), to be posted "if/when we have finished surveying people in all US counties". Because we targeted fewer total respondents than the total number of US counties, these posts will never be published. This formulation is therefore misleading, even if it is not explicitly deceptive. Given our desire to avoid leading respondents to publicly post political content (particularly xenophobic content, as in Experiment 3) as part of our survey, we and our Institutional Review Board felt comfortable with this formulation.

C.3 Considerations related to subject pool contamination (Experiments 1, 1R, and 3 and Auxiliary Experiments 2–3)

An important concern with deceptive or misleading experiments is that they can contaminate the subject pool by lowering trust in scientists and making respondents less likely to participate in future research studies. Of course, this can only happen if respondents know that they are being misled.

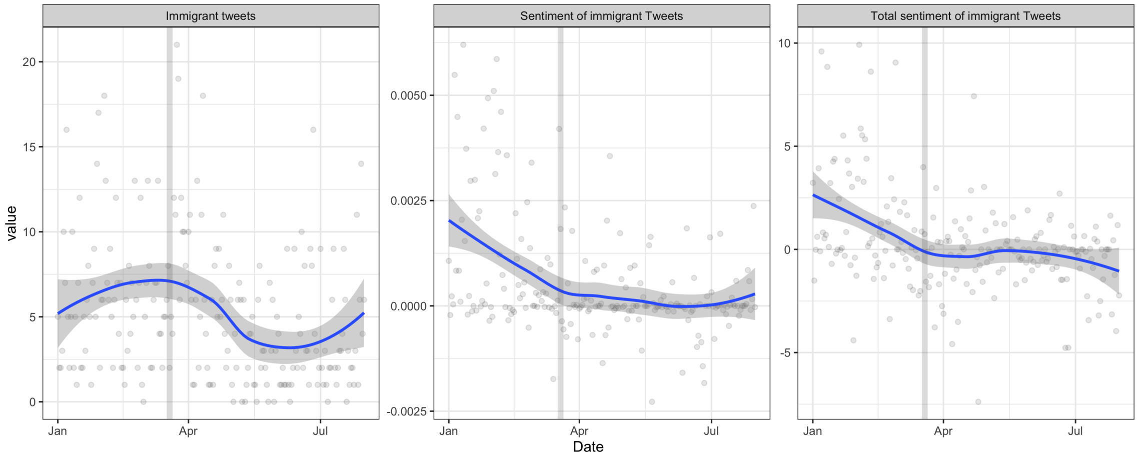
Subjects are told we will post their Tweets when and if we reach survey respondents on all US counties before August 1, 2021 (Experiment 3), December 1, 2021 (Experiment 1), March 1, 2022 (Auxiliary Experiment 2), and December 1, 2022 (Experiment 1R and Auxiliary Experiment 3). Although we privately targeted fewer respondents than the number of US counties, ensuring that this condition would not be met, subjects do not know (and never learn) this is the case. In other words, it is not possible for respondents to know that they have been misled about the implementation of the main outcomes (unless they independently find our working paper). Furthermore, concerns about contaminating the experimental subject pool are most important in an economic lab with clear rules against deception. In online survey marketplaces, where survey participants are expected to regularly participate in studies by psychologists in which explicit deception is common, considerations about contaminating the subject pool are less relevant.

C.4 Considerations related to starting political Twitter campaigns (Experiment 3)

As discussed in Appendix C.2, we designed our experiment to ensure that none of the Tweets would ever be posted. It is of course possible that respondents independently posted political content on Twitter as a result of our experiment. This is a concern for Experiment 3, in which respondents were exposed to a clip presenting a misleading narrative about the link between illegal immigration and crime.

To examine whether this was the case, we accessed all Twitter posts made by respondents between the date of experimental collection and August 1, 2021 (the date by which we promised respondents that our access to their accounts and any Twitter-related data would be deleted). We used simple text analysis techniques to identify which posts concern immigrants and quantify the sentiment and content of these posts. The results of this analysis are presented in Figure C.1 and Table C.1. We find no evidence that respondents in our experiment begin posting more immigrant-related Tweets or more negative content about immigrants after participating (Figure C.1). Restricting to the period after the experiment, we find no evidence that respondents in the *Cover* condition post more or fewer Tweets in general, more or fewer Tweets specifically about immigrants, or more or less negative Tweets about immigrants than respondents in our *No Cover* condition (Table C.1). This evidence further strengthens our confidence that our experiment did not contribute to anti-immigrant discourse on social media.

Figure C.1: Twitter activity of respondents before and after experiment



Notes: Figure C.1 presents various measures of the Twitter activity of respondents before and after Experiment 3, conducted between March 17 and March 22, 2021 (shaded in a gray rectangle). The left panel of the figure presents the average number of immigrant-related Tweets; the middle panel the average sentiment of immigrant-related Tweets; and the right panel the total expressed sentiment of immigrant-related Tweets.

Table C.1: Subsequent Twitter behavior of respondents

	Dependent variable:					
	Tw.	Tw. (w)	Imm. Tw.	Imm. Tw. (w)	Imm. sent.	Tot. imm. sent.
	(1)	(2)	(3)	(4)	(5)	(6)
Cover	−44.414	−9.298	−0.583	−0.152	0.005	0.024
	(29.941)	(9.462)	(0.416)	(0.117)	(0.012)	(0.062)
Constant	80.075***	35.951***	0.970***	0.383***	0.003	−0.052
	(20.862)	(6.593)	(0.290)	(0.082)	(0.008)	(0.043)
Observations	517	517	517	517	517	517

Notes: Table C.1 presents the results of our analysis of the subsequent Twitter behavior of the respondents in Experiment 3 between the end of our experiment and August 1, 2021. Table presents regressions of various measures of behavior on an indicator for whether the respondent was in the *Cover* condition: Columns 1 and 2 consider the total number of Tweets, Columns 3 and 4 the total number of immigrant-related Tweets, Column 5 the sentiment of immigrant-related Tweets, and Column 6 the sentiment of immigrant-related Tweets multiplied by the number of Tweets. Columns 2 and 4 winsorize the dependent variable at the 0.98 quantile.

D Additional Exhibits for Auxiliary Collections

This appendix reports balance and representativeness tables for all auxiliary collections.

Table D.1: Auxiliary Survey: Sample representativeness

	Defund	Pew (Inds, Dems and Reps)
Age	51.97	45.71
Black	0.10	0.12
White	0.83	0.74
Asian	0.04	0.04
Hispanic	0.06	0.15
Male	0.38	0.48
High school diploma	0.98	0.90
Bachelors degree or higher	0.47	0.31

Notes: Table displays mean characteristics, comparing the experimental sample with the 2018 Pew Research Center's American Trends Panel, Wave 39. Attriters are dropped from sample.

Table D.2: Auxiliary Experiment 1: Balance of covariates

	Overall		Article	No Article	*p*-value
	mean	std. dev.	mean	mean	(A = NA)
Age	36.91	14.06	37.24	36.59	0.46
Black	0.08	0.27	0.07	0.09	0.21
Asian	0.09	0.29	0.08	0.11	0.07
White	0.78	0.42	0.79	0.76	0.16
Hispanic	0.10	0.30	0.11	0.09	0.20
Male	0.49	0.50	0.49	0.50	0.70
High school diploma	1.00	0.06	0.99	1.00	0.04
Bachelors degree	0.63	0.48	0.62	0.65	0.36
Independent	0.21	0.41	0.21	0.21	0.90
Ideology	-1.09	0.80	-1.07	-1.11	0.43

Notes: *p*-values based on robust standard errors reported. Ideology is coded from -2 (very liberal) to 2 (very conservative).

Table D.3: Auxiliary Experiment 1: Sample representativeness

	Defund	Pew (Inds and Dems)
Age	36.91	45.86
Black	0.08	0.18
White	0.78	0.67
Asian	0.09	0.05
Hispanic	0.10	0.15
Male	0.49	0.46
High school diploma	1.00	0.89
Bachelors degree or higher	0.63	0.35

Notes: Table displays mean characteristics, comparing the experimental sample with the 2018 Pew Research Center's American Trends Panel, Wave 39. Attriters are dropped from sample.

Table D.4: Auxiliary Experiment 2: Balance of covariates

	Overall		Cover	No Cover	p-value
	mean	std. dev.	mean	mean	(C = NC)
Age	39.14	13.23	38.15	40.12	0.19
Black	0.14	0.34	0.17	0.10	0.07
Asian	0.04	0.21	0.03	0.06	0.10
White	0.77	0.42	0.74	0.80	0.22
Hispanic	0.16	0.37	0.17	0.15	0.74
Male	0.47	0.50	0.53	0.42	0.05
High school diploma	0.98	0.13	0.98	0.99	0.65
Bachelors degree	0.39	0.49	0.38	0.40	0.68
Independent	0.23	0.42	0.18	0.27	0.09
Republican	0.35	0.48	0.36	0.34	0.69
Democrat	0.43	0.50	0.46	0.40	0.28
Ideology	-0.19	1.19	-0.22	-0.16	0.66

Notes: p-values based on robust standard errors reported. Ideology is coded from -2 (very liberal) to 2 (very conservative).

Table D.5: Auxiliary Experiment 2: Sample representativeness

	Placebo	Pew (Inds, Dems and Reps)
Age	39.14	45.71
Black	0.14	0.12
White	0.77	0.74
Asian	0.04	0.04
Hispanic	0.16	0.15
Male	0.47	0.48
High school diploma	0.98	0.90
Bachelors degree or higher	0.39	0.31

Notes: Table displays mean characteristics, comparing the experimental sample with the 2018 Pew Research Center's American Trends Panel, Wave 39. Attriters are dropped from sample.

Table D.6: Auxiliary Experiment 3: Balance of covariates

	Overall		Cover	No Cover	*p*-value
	mean	std. dev.	mean	mean	(C = NC)
Age	45.79	15.36	45.13	46.45	0.33
Black	0.13	0.33	0.15	0.11	0.16
Asian	0.04	0.20	0.04	0.04	0.99
White	0.80	0.40	0.78	0.82	0.24
Hispanic	0.14	0.34	0.14	0.14	0.99
Male	0.53	0.50	0.54	0.52	0.63
High school diploma	0.99	0.12	0.99	0.98	0.70
Bachelors degree	0.49	0.50	0.48	0.49	0.76
Democrat	0.50	0.50	0.53	0.46	0.11
Ideology	-0.07	1.36	-0.16	0.02	0.13

Notes: *p*-values based on robust standard errors reported. Ideology is coded from -2 (very liberal) to 2 (very conservative).

Table D.7: Auxiliary Experiment 3: Sample representativeness

	Placebo	Pew (Dems and Reps)
Age	45.79	47.96
Black	0.13	0.15
White	0.80	0.73
Asian	0.04	0.04
Hispanic	0.14	0.13
Male	0.53	0.45
High school diploma	0.99	0.90
Bachelors degree or higher	0.49	0.33

Notes: Table displays mean characteristics, comparing the experimental sample with the 2018 Pew Research Center's American Trends Panel, Wave 39. Attriters are dropped from sample.

Table D.8: Auxiliary Experiment 4: Balance of covariates

	Overall		Cover	No Cover	*p*-value
	mean	std. dev.	mean	mean	(C = NC)
Age	34.79	14.33	34.62	34.98	0.78
Black	0.07	0.26	0.07	0.08	0.59
Asian	0.11	0.31	0.11	0.11	0.75
White	0.76	0.43	0.76	0.77	0.74
Hispanic	0.10	0.30	0.11	0.10	0.73
Male	0.46	0.50	0.47	0.45	0.62
High school diploma	0.99	0.08	0.99	1.00	0.09
Bachelors degree	0.48	0.50	0.45	0.52	0.10
Independent	0.35	0.48	0.37	0.34	0.62
Ideology	-0.79	0.86	-0.82	-0.77	0.53

Notes: *p*-values based on robust standard errors reported. Ideology is coded from -2 (very liberal) to 2 (very conservative).

Table D.9: Auxiliary Experiment 4: Sample representativeness

	Anticipated persuasion	Pew (Inds and Dems)
Age	34.79	45.86
Black	0.07	0.18
White	0.76	0.67
Asian	0.11	0.05
Hispanic	0.10	0.15
Male	0.46	0.46
High school diploma	0.99	0.89
Bachelors degree or higher	0.48	0.35

Notes: Table displays mean characteristics, comparing the experimental sample with the 2018 Pew Research Center's American Trends Panel, Wave 39. Attriters are dropped from sample.

Table D.10: Auxiliary Experiment 5: Balance of covariates

	Overall		Cover	No Cover	*p*-value
	mean	std. dev.	mean	mean	(C = NC)
Age	33.30	12.53	33.36	33.23	0.92
Black	0.08	0.27	0.08	0.07	0.87
Asian	0.13	0.34	0.14	0.12	0.69
White	0.75	0.43	0.75	0.75	0.95
Hispanic	0.12	0.32	0.14	0.10	0.17
Male	0.49	0.50	0.50	0.48	0.76
High school diploma	1.00	0.05	1.00	0.99	0.32
Bachelors degree	0.62	0.49	0.59	0.65	0.25
Independent	0.18	0.39	0.16	0.21	0.18
Ideology	-1.18	0.75	-1.16	-1.21	0.49

Notes: *p*-values based on robust standard errors reported. Ideology is coded from -2 (very liberal) to 2 (very conservative).

Table D.11: Auxiliary Experiment 5: Sample representativeness

	Motives	Pew (Inds and Dems)
Age	33.30	45.86
Black	0.08	0.18
White	0.75	0.67
Asian	0.13	0.05
Hispanic	0.12	0.15
Male	0.49	0.46
High school diploma	1.00	0.89
Bachelors degree or higher	0.62	0.35

Notes: Table displays mean characteristics, comparing the experimental sample with the 2018 Pew Research Center's American Trends Panel, Wave 39. Attriters are dropped from sample.

Table D.12: Auxiliary Experiment 6: Balance of covariates

	Overall		Cover	No Cover	*p*-value
	mean	std. dev.	mean	mean	(C = NC)
Panel A:			Higher-credibility		
Age	41.55	14.02	41.01	42.14	0.38
Black	0.21	0.41	0.20	0.22	0.53
Asian	0.05	0.21	0.05	0.04	0.52
White	0.70	0.46	0.70	0.69	0.95
Hispanic	0.15	0.36	0.15	0.15	0.86
Male	0.42	0.49	0.42	0.42	0.91
High school diploma	0.98	0.14	0.98	0.98	0.84
Bachelors degree	0.48	0.50	0.48	0.49	0.88
Independent	0.20	0.40	0.21	0.19	0.69
Ideology	-0.67	0.95	-0.68	-0.67	0.92
Panel B:			Lower-credibility		
Age	40.80	14.58	40.50	41.10	0.63
Black	0.19	0.40	0.20	0.19	0.79
Asian	0.04	0.19	0.03	0.04	0.63
White	0.71	0.45	0.74	0.69	0.26
Hispanic	0.17	0.37	0.15	0.18	0.45
Male	0.45	0.50	0.47	0.44	0.41
High school diploma	0.99	0.11	0.99	0.99	0.72
Bachelors degree	0.45	0.50	0.44	0.46	0.65
Independent	0.21	0.41	0.20	0.22	0.55
Ideology	-0.64	0.93	-0.69	-0.60	0.26

Notes: *p*-values based on robust standard errors reported. Ideology is coded from -2 (very liberal) to 2 (very conservative).

Table D.13: Auxiliary Experiment 6: Sample representativeness

	Defund	Pew (Inds and Dems)
Panel A:	Higher-credibility	
Age	41.55	45.86
Black	0.21	0.18
White	0.70	0.67
Asian	0.05	0.05
Hispanic	0.15	0.15
Male	0.42	0.46
High school diploma	0.98	0.89
Bachelors degree or higher	0.48	0.35
Panel B:	Lower-credibility	
Age	40.80	45.86
Black	0.19	0.18
White	0.71	0.67
Asian	0.04	0.05
Hispanic	0.17	0.15
Male	0.45	0.46
High school diploma	0.99	0.89
Bachelors degree or higher	0.45	0.35

Notes: Table displays mean characteristics, comparing the experimental sample with the 2018 Pew Research Center's American Trends Panel, Wave 39. Attriters are dropped from sample.

Table D.14: Auxiliary Experiment 6: lower-credibility variation, reweighted to match higher-credibility sample in Experiment 2 on demographics

	Auxiliary Experiment 6	
	(1)	(2)
Panel A:	*Belief partner donated*	
Cover	0.010	0.016
	(0.041)	(0.041)
No Cover mean	0.310	0.310
R^2	0.0001	0.042
Panel B:	*Denied bonus to partner*	
Cover	0.016	0.007
	(0.045)	(0.045)
No Cover mean	0.429	0.429
R^2	0.0003	0.058
Demographic controls	No	Yes
Observations	494	494

Notes: The dependent variable in Panel A is an indicator taking value 1 if the respondent reports believing that his or her matched partner donated to the US Border Crisis Children's Relief Fund. The dependent variable in Panel B is an indicator taking value 1 if the respondent denied his or her matched partner a $1 bonus. Columns 1–2 report results for the lower-credibility experiment. Demographic controls include age, age squared, a set of race indicators, a Hispanic indicator, a male indicator, and a set of education indicators. This table reweights observations to match the higher-credibility sample (Experiment 2) on observables. 12 observations are dropped as part of this reweighting. Robust standard errors are reported. *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively.

Table D.15: Auxiliary Experiment 7: Balance of covariates

	Overall		Cover	No Cover	p-value
	mean	std. dev.	mean	mean	(C = NC)
Age	35.37	14.58	35.27	35.46	0.89
Black	0.05	0.22	0.04	0.06	0.30
Asian	0.13	0.34	0.14	0.13	0.75
White	0.77	0.42	0.77	0.77	0.92
Hispanic	0.11	0.31	0.14	0.07	0.01
Male	0.50	0.50	0.50	0.49	0.79
High school diploma	1.00	0.06	0.99	1.00	0.16
Bachelors degree	0.60	0.49	0.60	0.59	0.88
Independent	0.22	0.41	0.24	0.19	0.19
Ideology	-1.08	0.83	-1.04	-1.13	0.21

Notes: p-values based on robust standard errors reported. Ideology is coded from -2 (very liberal) to 2 (very conservative).

Table D.16: Auxiliary Experiment 7: Sample representativeness

	Placebo	Pew (Inds and Dems)
Age	35.37	45.86
Black	0.05	0.18
White	0.77	0.67
Asian	0.13	0.05
Hispanic	0.11	0.15
Male	0.50	0.46
High school diploma	1.00	0.89
Bachelors degree or higher	0.60	0.35

Notes: Table displays mean characteristics, comparing the experimental sample with the 2018 Pew Research Center's American Trends Panel, Wave 39. Attriters are dropped from sample.

Table D.17: Auxiliary Experiment 8: Balance of covariates

	Overall		Article	No Article	p-value
	mean	std. dev.	mean	mean	(A = NA)
Age	44.59	15.45	44.80	44.39	0.56
Black	0.03	0.16	0.03	0.02	0.72
Asian	0.03	0.16	0.03	0.03	0.54
White	0.92	0.27	0.92	0.92	0.77
Hispanic	0.06	0.24	0.06	0.06	0.87
Male	0.43	0.50	0.42	0.44	0.38
High school diploma	0.98	0.13	0.98	0.99	0.07
Bachelors degree	0.41	0.49	0.40	0.42	0.45
Independent	0.19	0.39	0.19	0.19	0.99
Ideology	0.88	0.80	0.87	0.90	0.35

Notes: p-values based on robust standard errors reported. Ideology is coded from -2 (very liberal) to 2 (very conservative).

Table D.18: Auxiliary Experiment 8: Sample representativeness

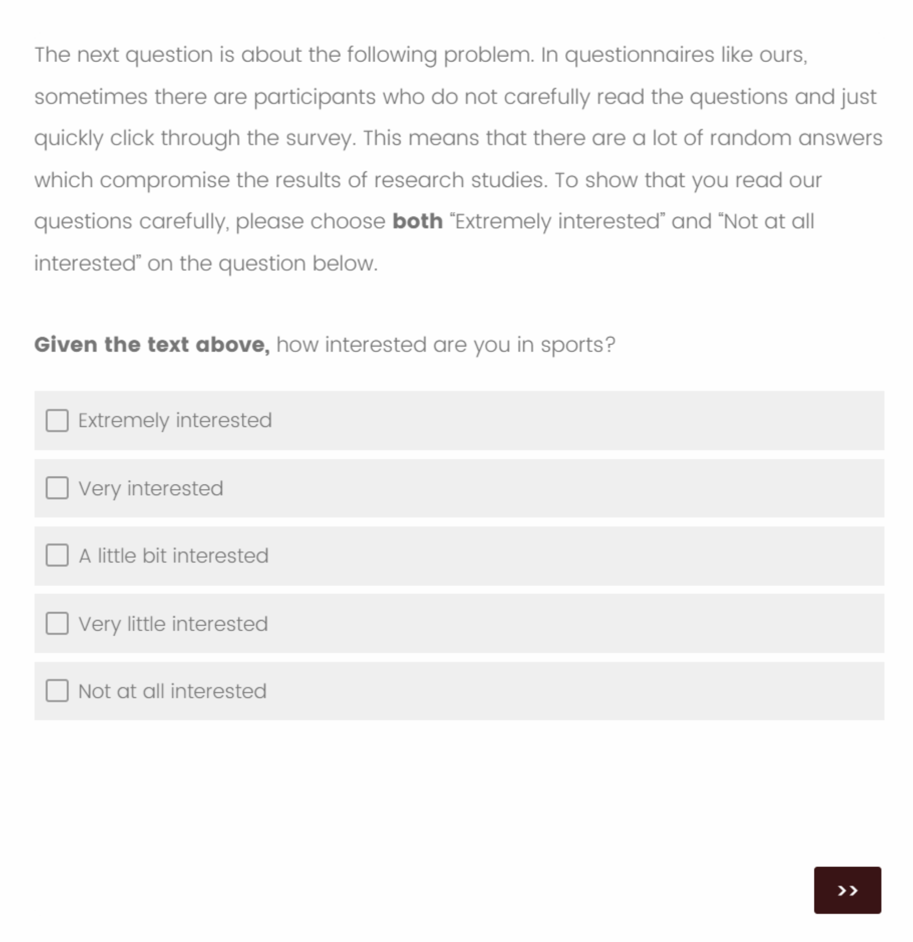
	Deport	Pew (Inds and Reps)
Age	44.59	47.17
Black	0.03	0.05
White	0.92	0.83
Asian	0.03	0.03
Hispanic	0.06	0.11
Male	0.43	0.52
High school diploma	0.98	0.93
Bachelors degree or higher	0.41	0.31

Notes: Table displays mean characteristics, comparing the experimental sample with the 2018 Pew Research Center's American Trends Panel, Wave 39. Attriters are dropped from sample.

E Experimental Instructions

E.1 Experiment 1 and 1R: Expression of dissent — Democrats

E.1.1 Attention screener

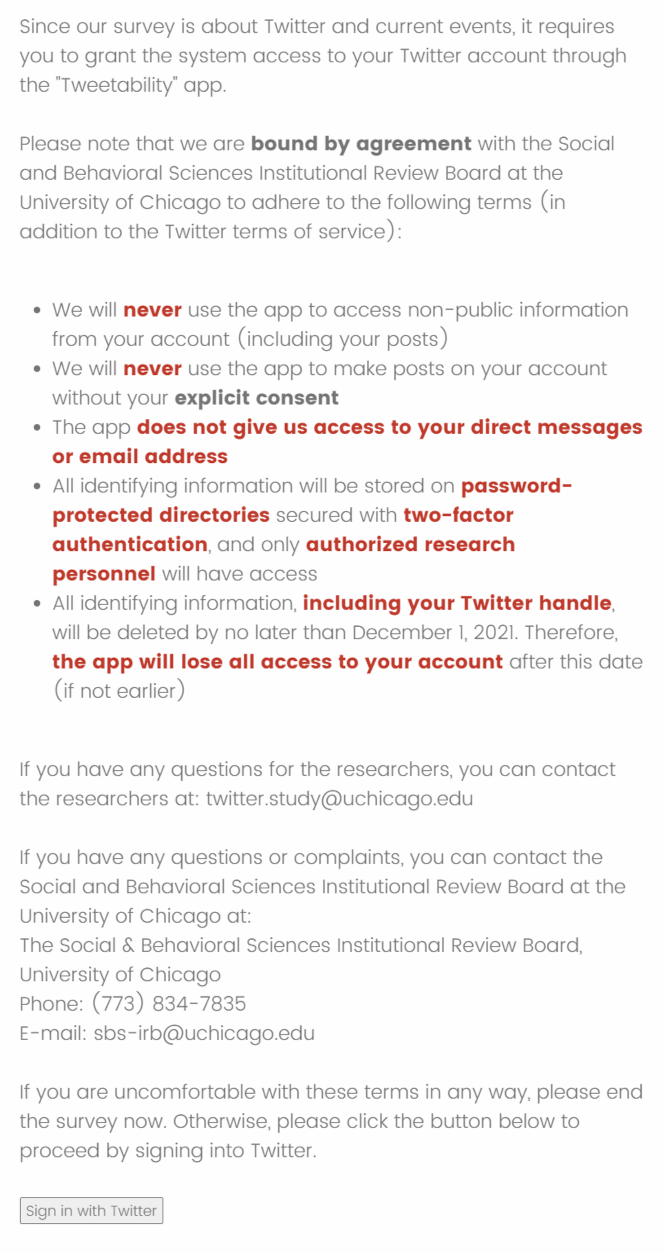


The next question is about the following problem. In questionnaires like ours, sometimes there are participants who do not carefully read the questions and just quickly click through the survey. This means that there are a lot of random answers which compromise the results of research studies. To show that you read our questions carefully, please choose **both** "Extremely interested" and "Not at all interested" on the question below.

Given the text above, how interested are you in sports?

- □ Extremely interested
- □ Very interested
- □ A little bit interested
- □ Very little interested
- □ Not at all interested

E.1.2 Twitter information and login



Since our survey is about Twitter and current events, it requires you to grant the system access to your Twitter account through the "Tweetability" app.

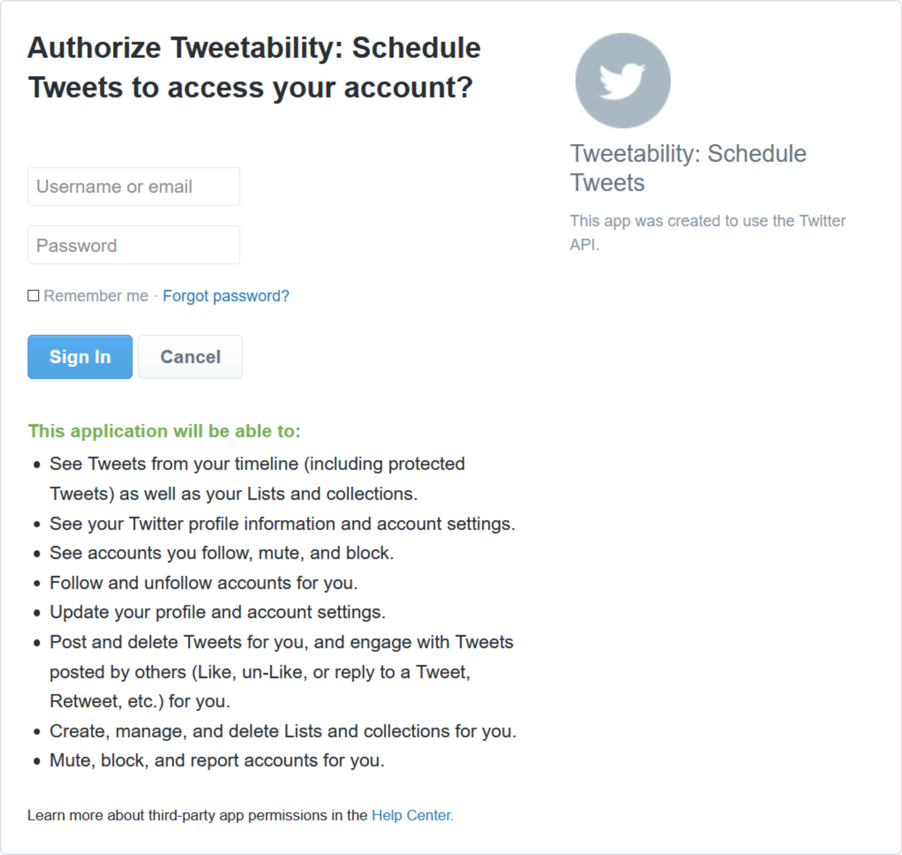
Please note that we are **bound by agreement** with the Social and Behavioral Sciences Institutional Review Board at the University of Chicago to adhere to the following terms (in addition to the Twitter terms of service):

- We will **never** use the app to access non-public information from your account (including your posts)
- We will **never** use the app to make posts on your account without your **explicit consent**
- The app **does not give us access to your direct messages or email address**
- All identifying information will be stored on **password-protected directories** secured with **two-factor authentication**, and only **authorized research personnel** will have access
- All identifying information, **including your Twitter handle**, will be deleted by no later than December 1, 2021. Therefore, **the app will lose all access to your account** after this date (if not earlier)

If you have any questions for the researchers, you can contact the researchers at: twitter.study@uchicago.edu

If you have any questions or complaints, you can contact the Social and Behavioral Sciences Institutional Review Board at the University of Chicago at:
The Social & Behavioral Sciences Institutional Review Board, University of Chicago
Phone: (773) 834-7835
E-mail: sbs-irb@uchicago.edu

If you are uncomfortable with these terms in any way, please end the survey now. Otherwise, please click the button below to proceed by signing into Twitter.



Authorize Tweetability: Schedule Tweets to access your account?

Tweetability: Schedule Tweets

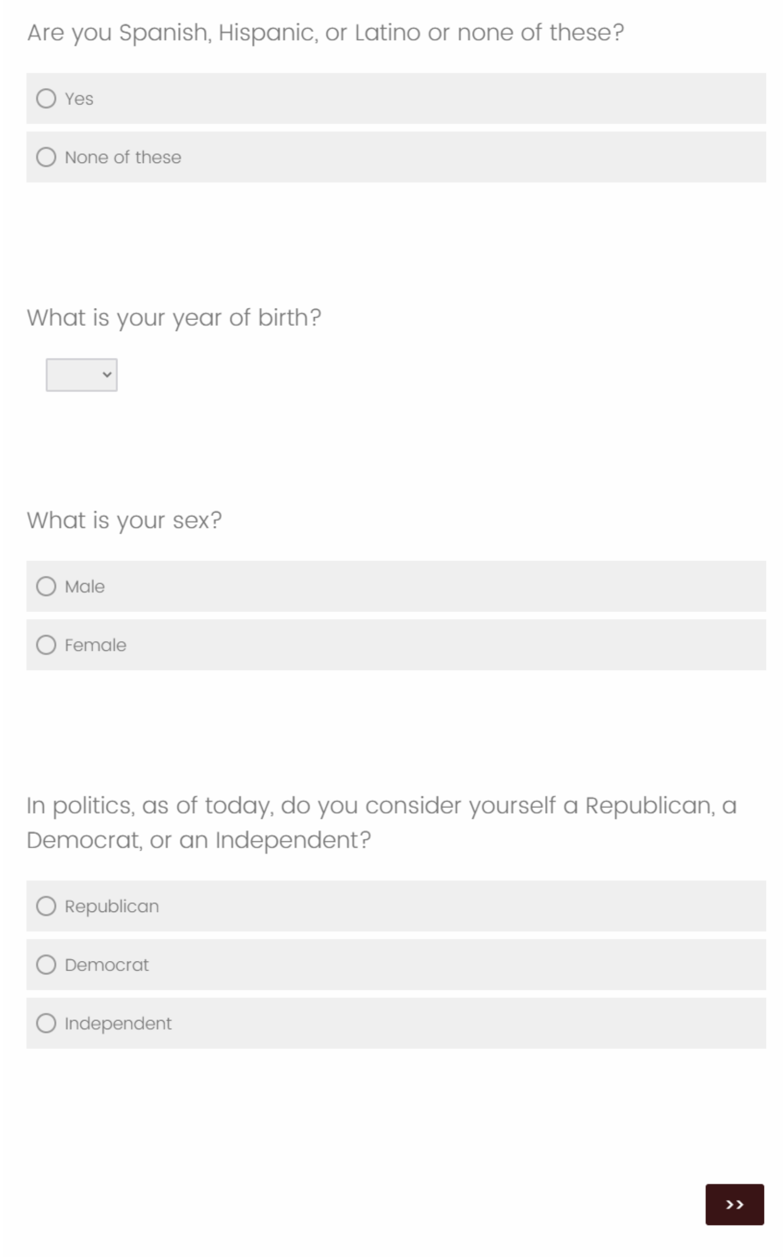
This app was created to use the Twitter API.

This application will be able to:

- See Tweets from your timeline (including protected Tweets) as well as your Lists and collections.
- See your Twitter profile information and account settings.
- See accounts you follow, mute, and block.
- Follow and unfollow accounts for you.
- Update your profile and account settings.
- Post and delete Tweets for you, and engage with Tweets posted by others (Like, un-Like, or reply to a Tweet, Retweet, etc.) for you.
- Create, manage, and delete Lists and collections for you.
- Mute, block, and report accounts for you.

Learn more about third-party app permissions in the Help Center.

E.1.3 Background questions



Are you Spanish, Hispanic, or Latino or none of these?

○ Yes

○ None of these

What is your year of birth?

What is your sex?

○ Male

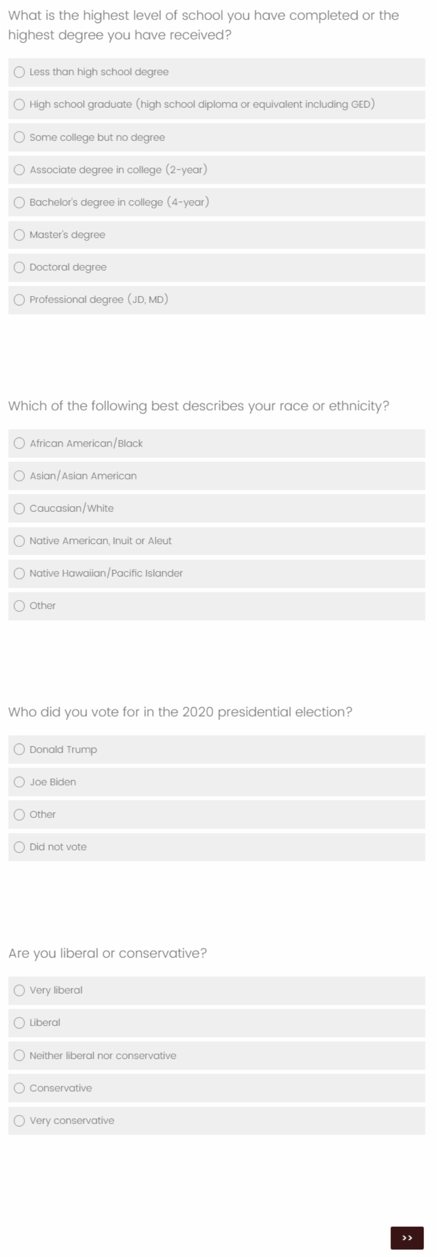
○ Female

In politics, as of today, do you consider yourself a Republican, a Democrat, or an Independent?

○ Republican

○ Democrat

○ Independent



What is the highest level of school you have completed or the highest degree you have received?

○ Less than high school degree

○ High school graduate (high school diploma or equivalent including GED)

○ Some college but no degree

○ Associate degree in college (2-year)

○ Bachelor's degree in college (4-year)

○ Master's degree

○ Doctoral degree

○ Professional degree (JD, MD)

Which of the following best describes your race or ethnicity?

○ African American/Black

○ Asian/Asian American

○ Caucasian/White

○ Native American, Inuit or Aleut

○ Native Hawaiian/Pacific Islander

○ Other

Who did you vote for in the 2020 presidential election?

○ Donald Trump

○ Joe Biden

○ Other

○ Did not vote

Are you liberal or conservative?

○ Very liberal

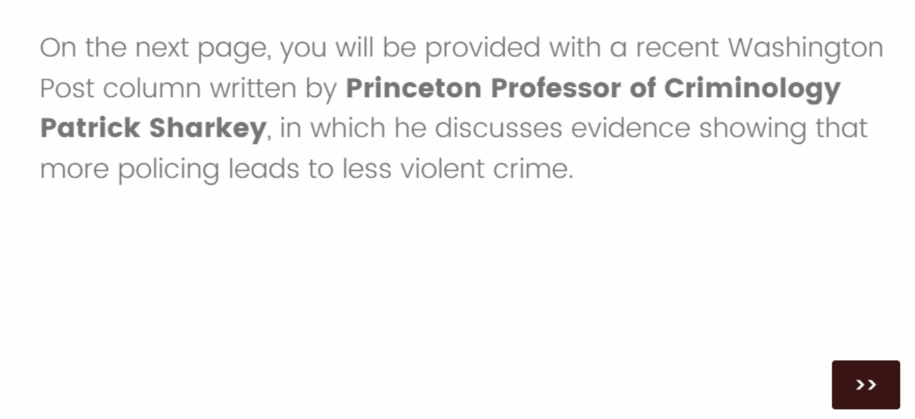
○ Liberal

○ Neither liberal nor conservative

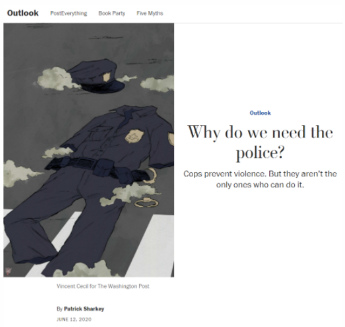
○ Conservative

○ Very conservative

E.1.4 Pre-treatment outcomes



On the next page, you will be provided with a recent Washington Post column written by **Princeton Professor of Criminology Patrick Sharkey**, in which he discusses evidence showing that more policing leads to less violent crime.



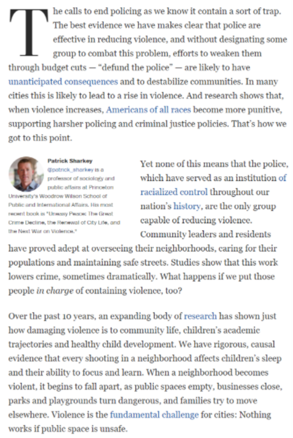
Outlook

Why do we need the police?

Cops prevent violence. But they aren't the only ones who can do it.

Vincent Cecil for The Washington Post

By **Patrick Sharkey**
JUNE 12, 2020

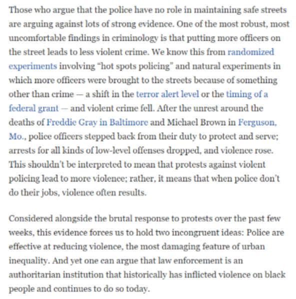


The calls to end policing as we know it contain a sort of trap. The best evidence we have makes clear that police are effective in reducing violence, and without designating some group to combat this problem, efforts to weaken them through budget cuts — "defund the police" — are likely to have unanticipated consequences and to destabilize communities. In many cities this is likely to lead to a rise in violence. And research shows that, when violence increases, Americans of all races become more punitive, supporting harsher policing and criminal justice policies. That's how we got to this point.

Patrick Sharkey @patrick_sharkey is a professor of sociology and public affairs at Princeton University's Woodrow Wilson School of Public and International Affairs. His most recent book is "Uneasy Peace: The Great Crime Decline, the Renewal of City Life, and the Next War on Violence."

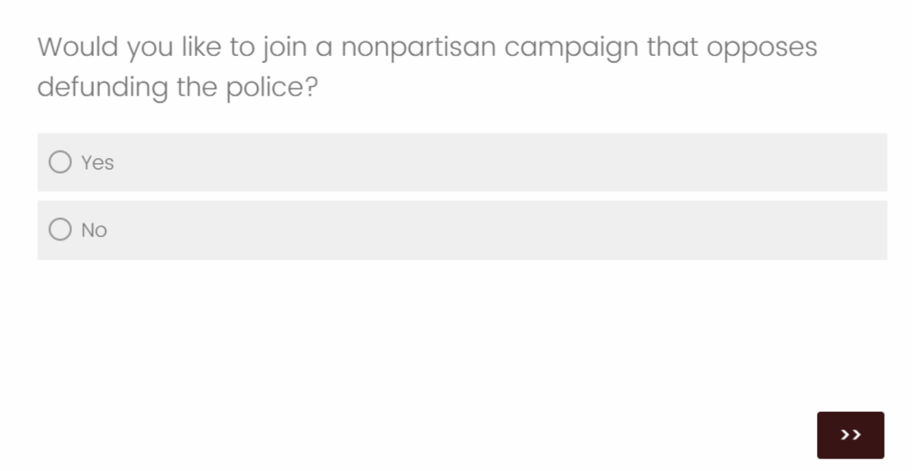
Yet none of this means that the police, which have served as an institution of racialized control throughout our nation's history, are the only group capable of reducing violence. Community leaders and residents have proved adept at overseeing their neighborhoods, caring for their populations and maintaining safe streets. Studies show that this work lowers crime, sometimes dramatically. What happens if we put those people *in charge* of containing violence, too?

Over the past 10 years, an expanding body of research has shown just how damaging violence is to community life, children's academic trajectories and healthy child development. We have rigorous, causal evidence that every shooting in a neighborhood affects children's sleep and their ability to focus and learn. When a neighborhood becomes violent, it begins to fall apart, as public spaces empty, businesses close, parks and playgrounds turn dangerous, and families try to move elsewhere. Violence is the fundamental challenge for cities: Nothing works if public space is unsafe.



Those who argue that the police have no role in maintaining safe streets are arguing against lots of strong evidence. One of the most robust, most uncomfortable findings in criminology is that putting more officers on the street leads to less violent crime. We know this from randomized experiments involving "hot spots policing" and natural experiments in which more officers were brought to the streets because of something other than crime — a shift in the terror alert level or the timing of a federal grant — and violent crime fell. After the unrest around the deaths of Freddie Gray in Baltimore and Michael Brown in Ferguson, Mo., police officers stepped back from their duty to protect and serve; arrests for all kinds of low-level offenses dropped, and violence rose. This shouldn't be interpreted to mean that protests against violent policing lead to more violence; rather, it means that when police don't do their jobs, violence often results.

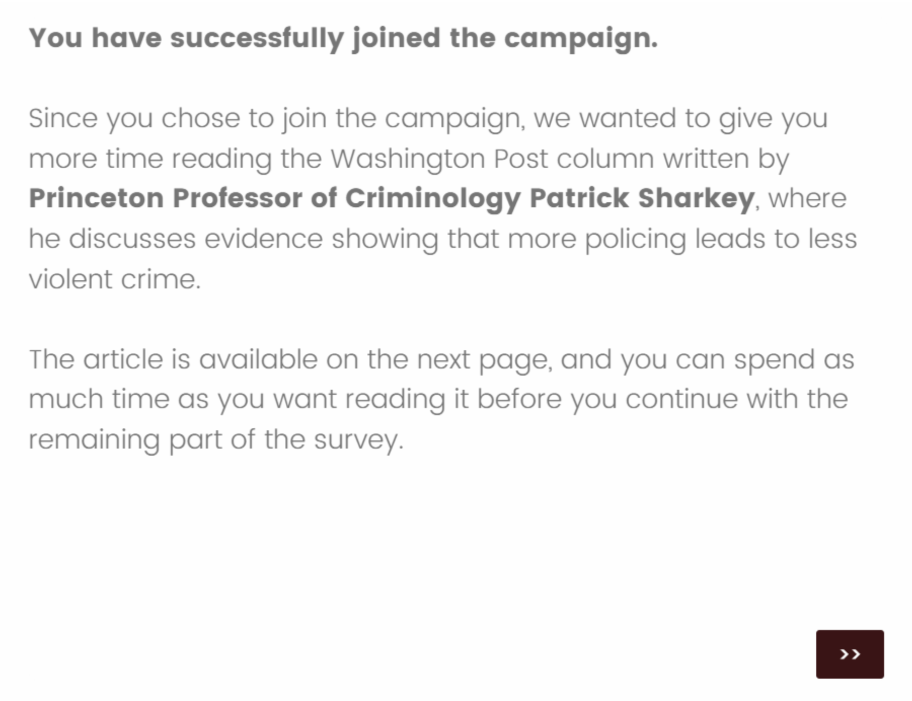
Considered alongside the brutal response to protests over the past few weeks, this evidence forces us to hold two incongruent ideas: Police are effective at reducing violence, the most damaging feature of urban inequality. And yet one can argue that law enforcement is an authoritarian institution that historically has inflicted violence on black people and continues to do so today.



Would you like to join a nonpartisan campaign that opposes defunding the police?

○ Yes

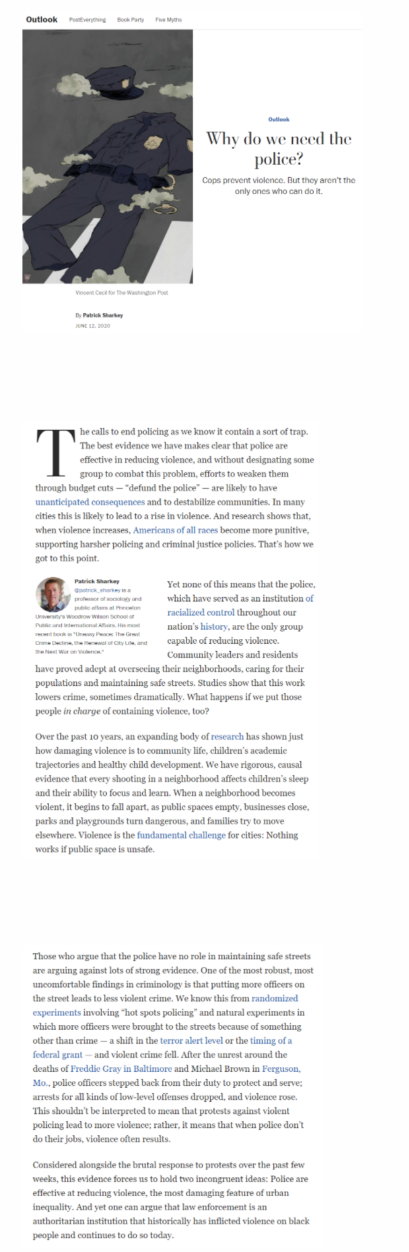
○ No



You have successfully joined the campaign.

Since you chose to join the campaign, we wanted to give you more time reading the Washington Post column written by **Princeton Professor of Criminology Patrick Sharkey**, where he discusses evidence showing that more policing leads to less violent crime.

The article is available on the next page, and you can spend as much time as you want reading it before you continue with the remaining part of the survey.



Outlook

Why do we need the police?

Cops prevent violence. But they aren't the only ones who can do it.

Vincent Cecil for The Washington Post

By **Patrick Sharkey**

JUNE 12, 2020

The calls to end policing as we know it contain a sort of trap. The best evidence we have makes clear that police are effective in reducing violence, and without designating some group to combat this problem, efforts to weaken them through budget cuts — "defund the police" — are likely to have unanticipated consequences and to destabilize communities. In many cities this is likely to lead to a rise in violence. And research shows that, when violence increases, Americans of all races become more punitive, supporting harsher policing and criminal justice policies. That's how we got to this point.

Patrick Sharkey @patrick_sharkey is a professor of sociology and public affairs at Princeton University's Woodrow Wilson School of Public and International Affairs. His most recent book is "Uneasy Peace: The Great Crime Decline, the Renewal of City Life, and the Next War on Violence."

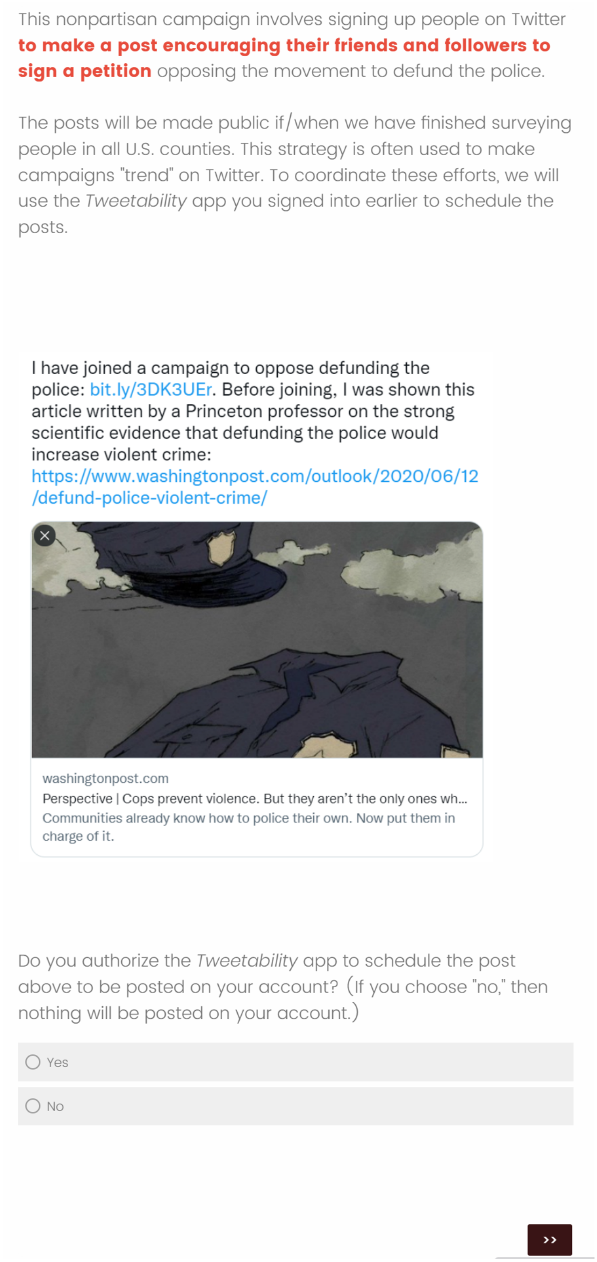
Yet none of this means that the police, which have served as an institution of racialized control throughout our nation's history, are the only group capable of reducing violence. Community leaders and residents have proved adept at overseeing their neighborhoods, caring for their populations and maintaining safe streets. Studies show that this work lowers crime, sometimes dramatically. What happens if we put those people *in charge* of containing violence, too?

Over the past 10 years, an expanding body of research has shown just how damaging violence is to community life, children's academic trajectories and healthy child development. We have rigorous, causal evidence that every shooting in a neighborhood affects children's sleep and their ability to focus and learn. When a neighborhood becomes violent, it begins to fall apart, as public spaces empty, businesses close, parks and playgrounds turn dangerous, and families try to move elsewhere. Violence is the fundamental challenge for cities: Nothing works if public space is unsafe.

Those who argue that the police have no role in maintaining safe streets are arguing against lots of strong evidence. One of the most robust, most uncomfortable findings in criminology is that putting more officers on the street leads to less violent crime. We know this from randomized experiments involving "hot spots policing" and natural experiments in which more officers were brought to the streets because of something other than crime — a shift in the terror alert level or the timing of a federal grant — and violent crime fell. After the unrest around the deaths of Freddie Gray in Baltimore and Michael Brown in Ferguson, Mo., police officers stepped back from their duty to protect and serve; arrests for all kinds of low-level offenses dropped, and violence rose. This shouldn't be interpreted to mean that protests against violent policing lead to more violence; rather, it means that when police don't do their jobs, violence often results.

Considered alongside the brutal response to protests over the past few weeks, this evidence forces us to hold two incongruent ideas: Police are effective at reducing violence, the most damaging feature of urban inequality. And yet one can argue that law enforcement is an authoritarian institution that historically has inflicted violence on black people and continues to do so today.

E.1.5 Treatment: “Before” wording (rationale)



This nonpartisan campaign involves signing up people on Twitter **to make a post encouraging their friends and followers to sign a petition** opposing the movement to defund the police.

The posts will be made public if/when we have finished surveying people in all U.S. counties. This strategy is often used to make campaigns "trend" on Twitter. To coordinate these efforts, we will use the *Tweetability* app you signed into earlier to schedule the posts.

I have joined a campaign to oppose defunding the police: bit.ly/3DK3UEr. Before joining, I was shown this article written by a Princeton professor on the strong scientific evidence that defunding the police would increase violent crime: https://www.washingtonpost.com/outlook/2020/06/12/defund-police-violent-crime/

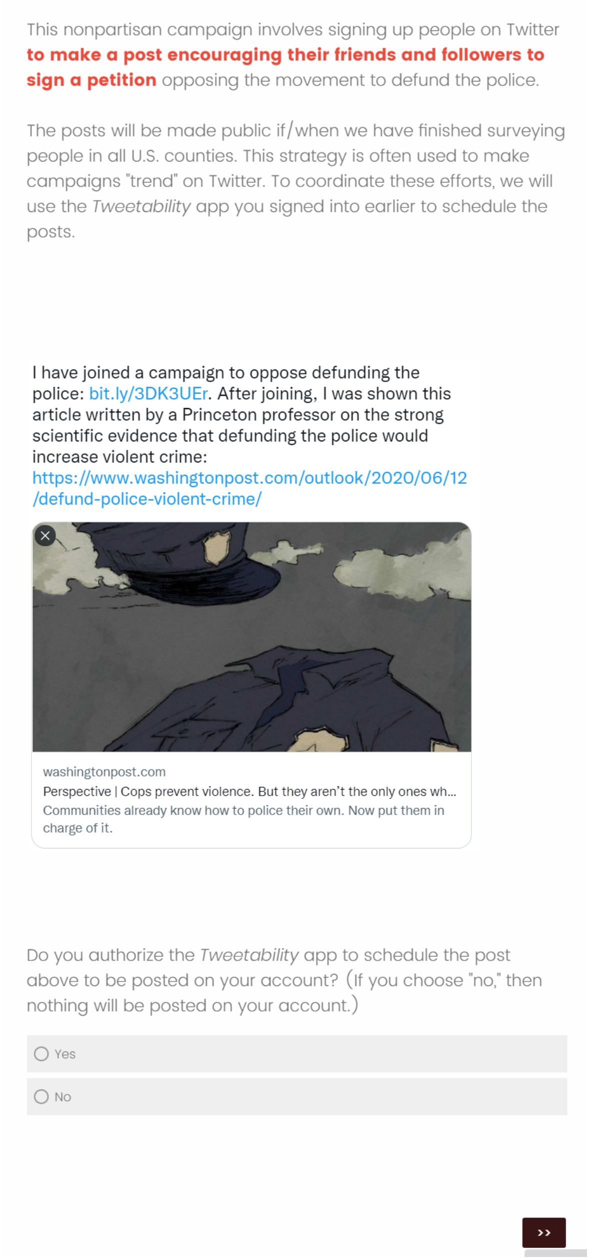
washingtonpost.com
Perspective | Cops prevent violence. But they aren’t the only ones wh...
Communities already know how to police their own. Now put them in charge of it.

Do you authorize the *Tweetability* app to schedule the post above to be posted on your account? (If you choose "no," then nothing will be posted on your account.)

○ Yes

○ No

E.1.6 Treatment: "After" wording (no rationale)



This nonpartisan campaign involves signing up people on Twitter **to make a post encouraging their friends and followers to sign a petition** opposing the movement to defund the police.

The posts will be made public if/when we have finished surveying people in all U.S. counties. This strategy is often used to make campaigns "trend" on Twitter. To coordinate these efforts, we will use the *Tweetability* app you signed into earlier to schedule the posts.

I have joined a campaign to oppose defunding the police: bit.ly/3DK3UEr. After joining, I was shown this article written by a Princeton professor on the strong scientific evidence that defunding the police would increase violent crime: https://www.washingtonpost.com/outlook/2020/06/12/defund-police-violent-crime/

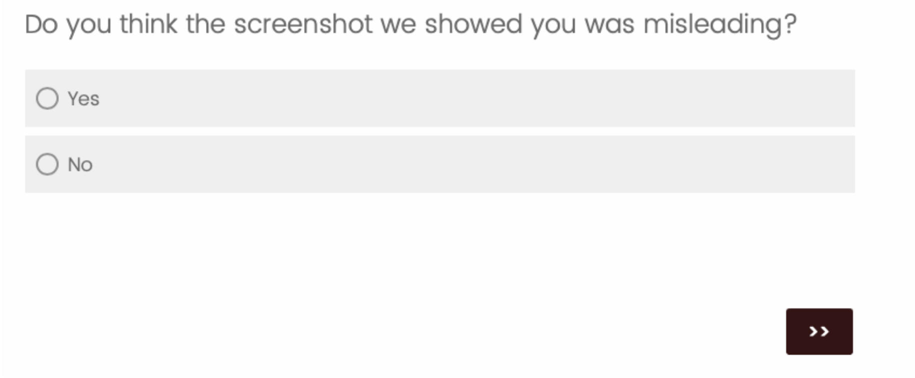
washingtonpost.com
Perspective | Cops prevent violence. But they aren't the only ones wh...
Communities already know how to police their own. Now put them in charge of it.

Do you authorize the *Tweetability* app to schedule the post above to be posted on your account? (If you choose "no," then nothing will be posted on your account.)

○ Yes

○ No

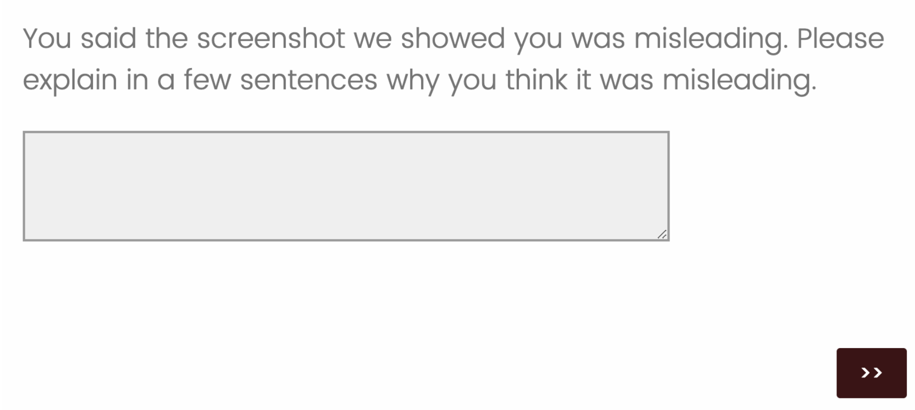
E.1.7 Post-treatment questions (Experiment 1R only)



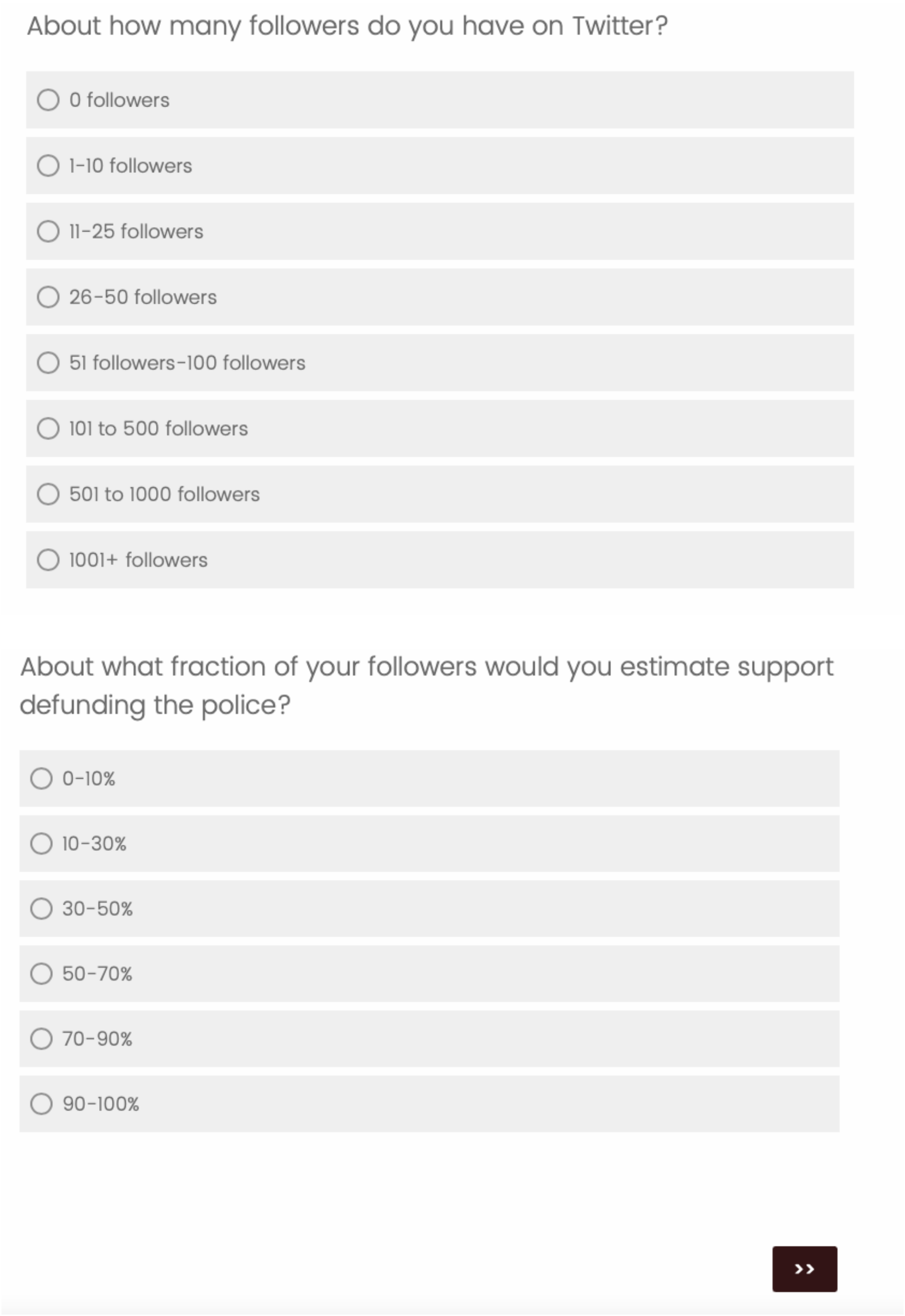
Do you think the screenshot we showed you was misleading?

○ Yes

○ No



You said the screenshot we showed you was misleading. Please explain in a few sentences why you think it was misleading.



About how many followers do you have on Twitter?

○ 0 followers

○ 1-10 followers

○ 11-25 followers

○ 26-50 followers

○ 51 followers-100 followers

○ 101 to 500 followers

○ 501 to 1000 followers

○ 1001+ followers

About what fraction of your followers would you estimate support defunding the police?

○ 0-10%

○ 10-30%

○ 30-50%

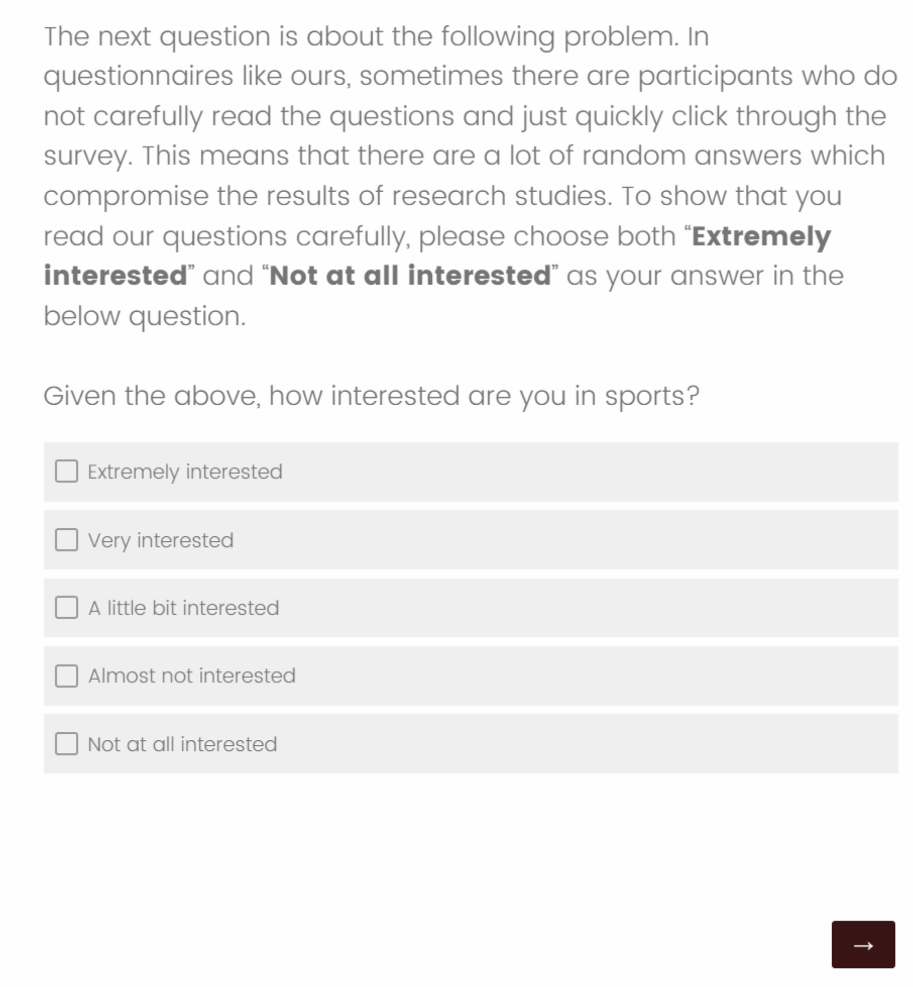
○ 50-70%

○ 70-90%

○ 90-100%

E.2 Experiment 2: Interpretation of dissent – Democrats

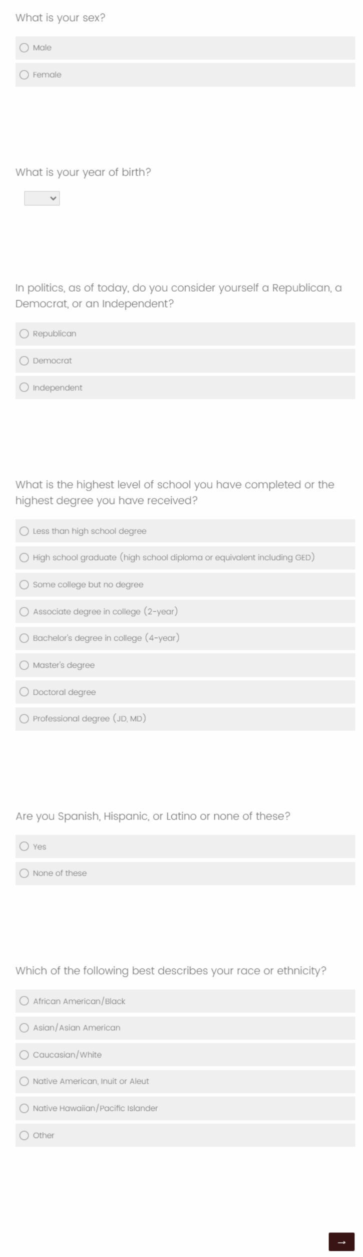
E.2.1 Attention screener and background questions



The next question is about the following problem. In questionnaires like ours, sometimes there are participants who do not carefully read the questions and just quickly click through the survey. This means that there are a lot of random answers which compromise the results of research studies. To show that you read our questions carefully, please choose both "**Extremely interested**" and "**Not at all interested**" as your answer in the below question.

Given the above, how interested are you in sports?

- □ Extremely interested
- □ Very interested
- □ A little bit interested
- □ Almost not interested
- □ Not at all interested



What is your sex?

○ Male

○ Female

What is your year of birth?

In politics, as of today, do you consider yourself a Republican, a Democrat, or an Independent?

○ Republican

○ Democrat

○ Independent

What is the highest level of school you have completed or the highest degree you have received?

○ Less than high school degree

○ High school graduate (high school diploma or equivalent including GED)

○ Some college but no degree

○ Associate degree in college (2-year)

○ Bachelor's degree in college (4-year)

○ Master's degree

○ Doctoral degree

○ Professional degree (JD, MD)

Are you Spanish, Hispanic, or Latino or none of these?

○ Yes

○ None of these

Which of the following best describes your race or ethnicity?

○ African American/Black

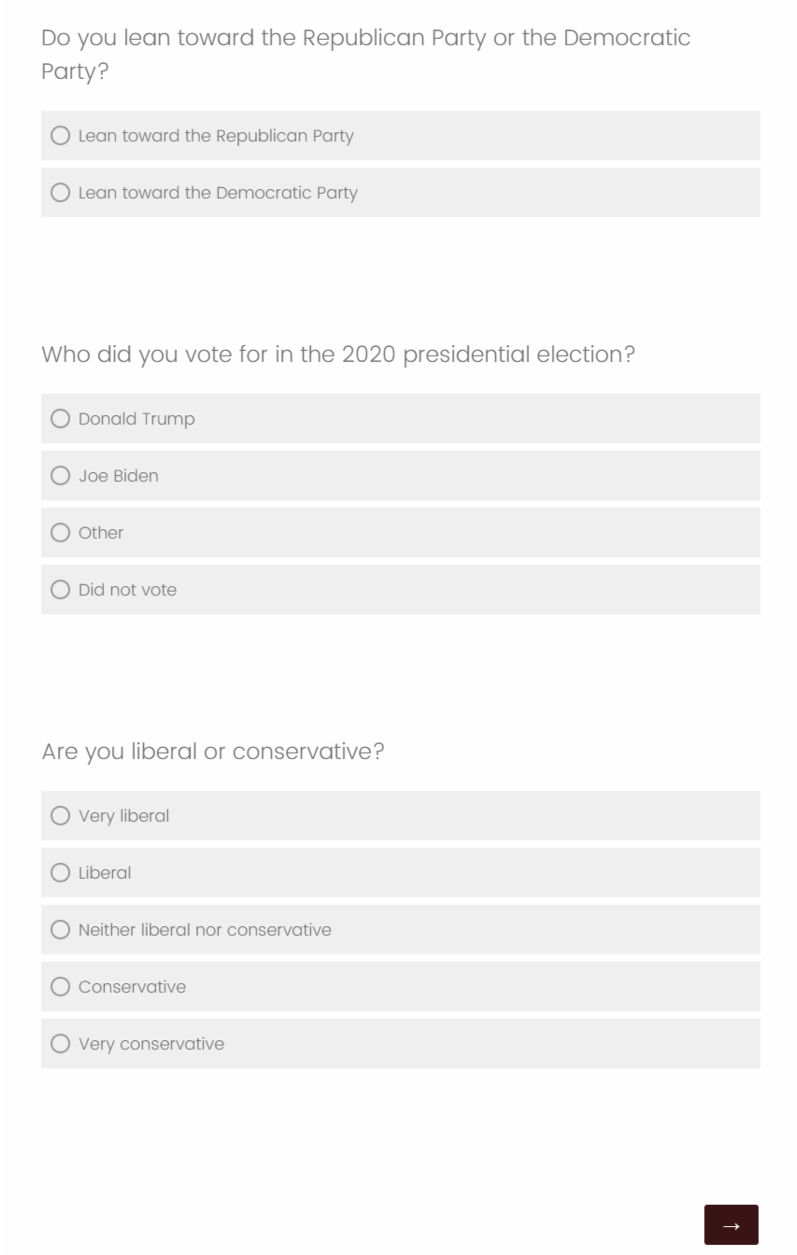
○ Asian/Asian American

○ Caucasian/White

○ Native American, Inuit or Aleut

○ Native Hawaiian/Pacific Islander

○ Other



Do you lean toward the Republican Party or the Democratic Party?

○ Lean toward the Republican Party

○ Lean toward the Democratic Party

Who did you vote for in the 2020 presidential election?

○ Donald Trump

○ Joe Biden

○ Other

○ Did not vote

Are you liberal or conservative?

○ Very liberal

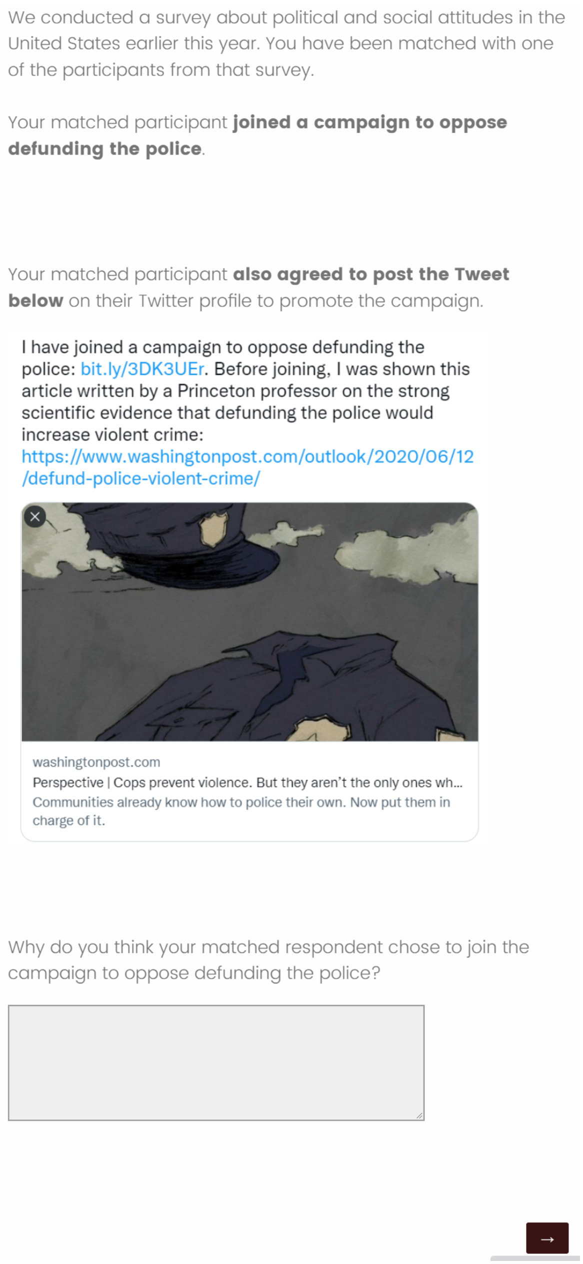
○ Liberal

○ Neither liberal nor conservative

○ Conservative

○ Very conservative

E.2.2 Treatment: “Before” wording (rationale)



We conducted a survey about political and social attitudes in the United States earlier this year. You have been matched with one of the participants from that survey.

Your matched participant **joined a campaign to oppose defunding the police**.

Your matched participant **also agreed to post the Tweet below** on their Twitter profile to promote the campaign.

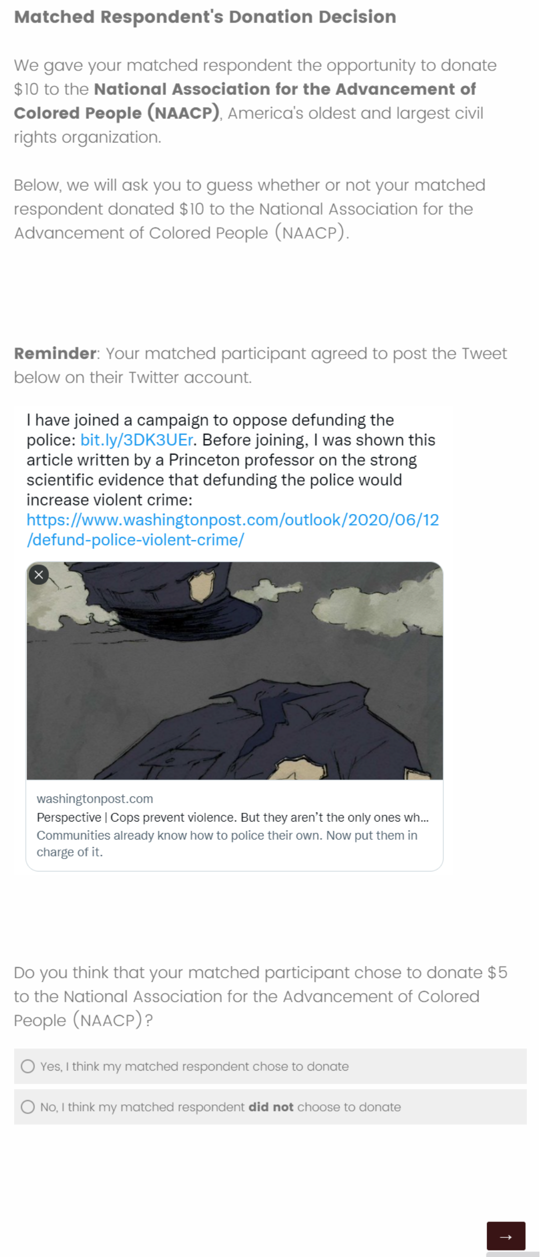
I have joined a campaign to oppose defunding the police: bit.ly/3DK3UEr. Before joining, I was shown this article written by a Princeton professor on the strong scientific evidence that defunding the police would increase violent crime: https://www.washingtonpost.com/outlook/2020/06/12/defund-police-violent-crime/

washingtonpost.com

Perspective | Cops prevent violence. But they aren’t the only ones wh...

Communities already know how to police their own. Now put them in charge of it.

Why do you think your matched respondent chose to join the campaign to oppose defunding the police?



Matched Respondent's Donation Decision

We gave your matched respondent the opportunity to donate $10 to the **National Association for the Advancement of Colored People (NAACP)**, America's oldest and largest civil rights organization.

Below, we will ask you to guess whether or not your matched respondent donated $10 to the National Association for the Advancement of Colored People (NAACP).

Reminder: Your matched participant agreed to post the Tweet below on their Twitter account.

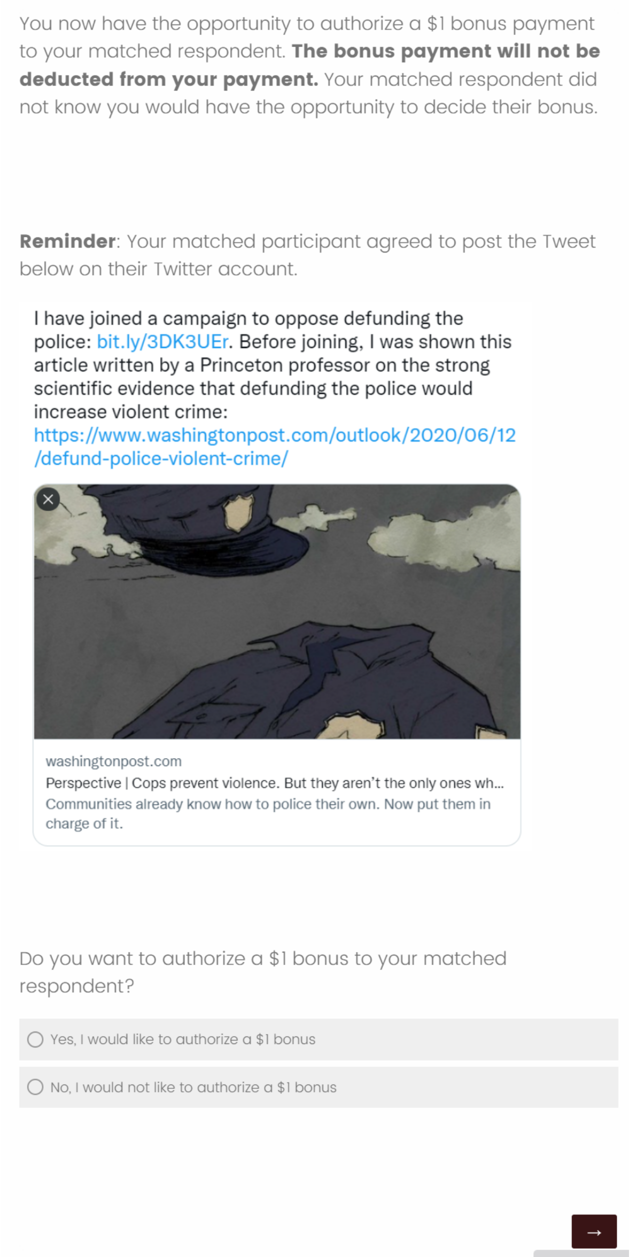
I have joined a campaign to oppose defunding the police: bit.ly/3DK3UEr. Before joining, I was shown this article written by a Princeton professor on the strong scientific evidence that defunding the police would increase violent crime: https://www.washingtonpost.com/outlook/2020/06/12/defund-police-violent-crime/

washingtonpost.com
Perspective | Cops prevent violence. But they aren't the only ones wh...
Communities already know how to police their own. Now put them in charge of it.

Do you think that your matched participant chose to donate $5 to the National Association for the Advancement of Colored People (NAACP)?

○ Yes, I think my matched respondent chose to donate

○ No, I think my matched respondent **did not** choose to donate



You now have the opportunity to authorize a $1 bonus payment to your matched respondent. **The bonus payment will not be deducted from your payment.** Your matched respondent did not know you would have the opportunity to decide their bonus.

Reminder: Your matched participant agreed to post the Tweet below on their Twitter account.

I have joined a campaign to oppose defunding the police: bit.ly/3DK3UEr. Before joining, I was shown this article written by a Princeton professor on the strong scientific evidence that defunding the police would increase violent crime: https://www.washingtonpost.com/outlook/2020/06/12/defund-police-violent-crime/

washingtonpost.com

Perspective | Cops prevent violence. But they aren't the only ones wh...

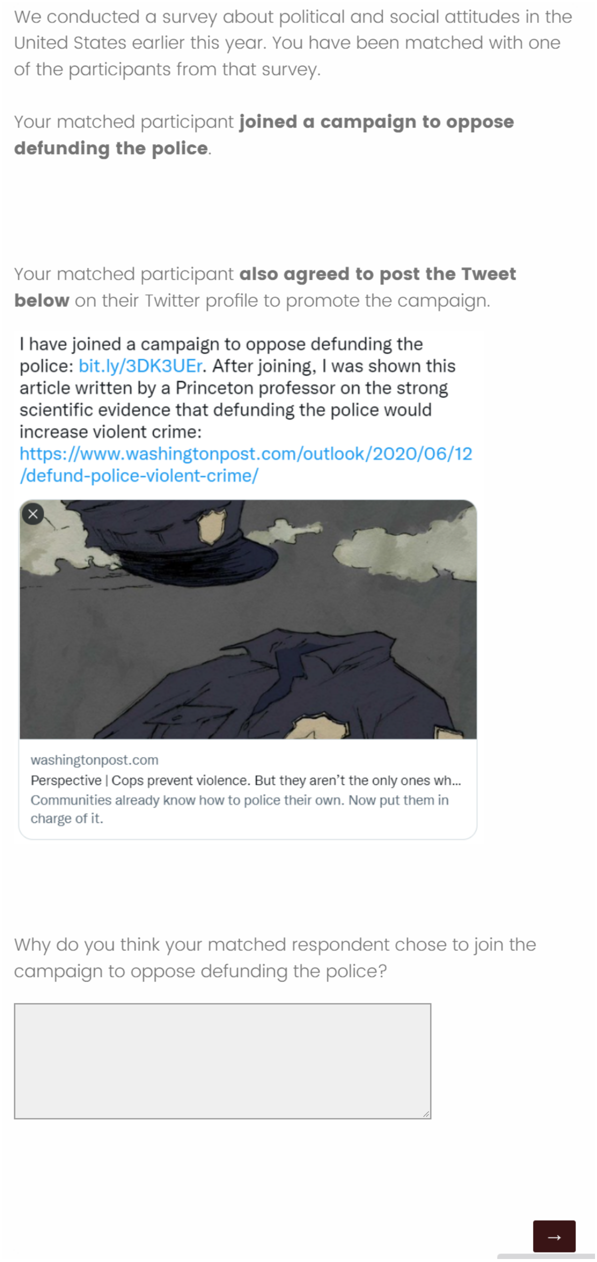
Communities already know how to police their own. Now put them in charge of it.

Do you want to authorize a $1 bonus to your matched respondent?

○ Yes, I would like to authorize a $1 bonus

○ No, I would not like to authorize a $1 bonus

E.2.3 Treatment: “After” wording (no rationale)



We conducted a survey about political and social attitudes in the United States earlier this year. You have been matched with one of the participants from that survey.

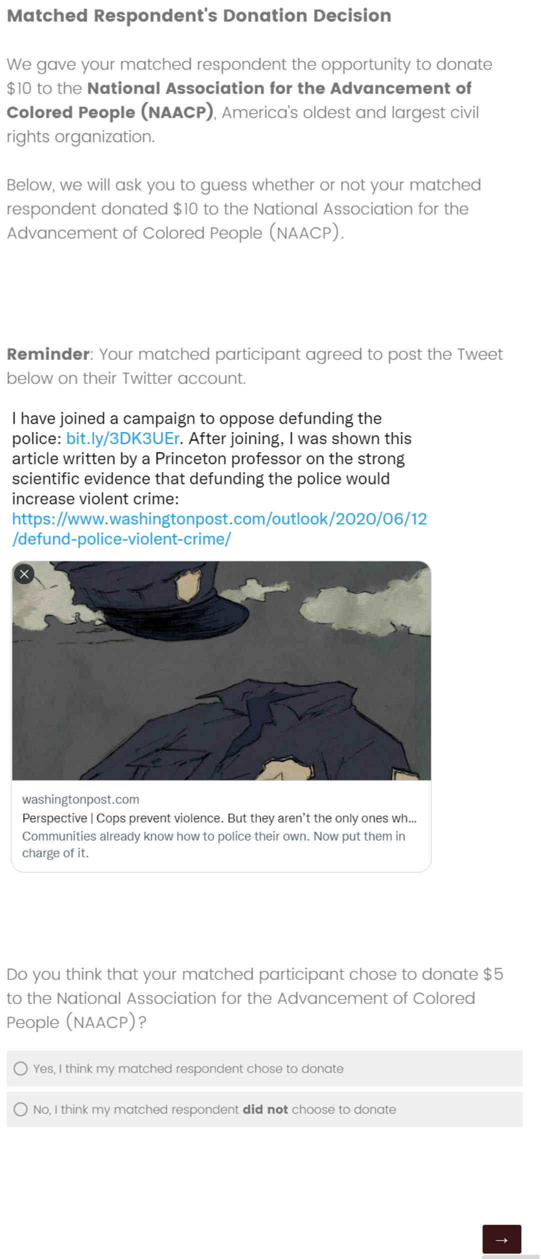
Your matched participant **joined a campaign to oppose defunding the police**.

Your matched participant **also agreed to post the Tweet below** on their Twitter profile to promote the campaign.

I have joined a campaign to oppose defunding the police: bit.ly/3DK3UEr. After joining, I was shown this article written by a Princeton professor on the strong scientific evidence that defunding the police would increase violent crime: https://www.washingtonpost.com/outlook/2020/06/12/defund-police-violent-crime/

washingtonpost.com
Perspective | Cops prevent violence. But they aren’t the only ones wh...
Communities already know how to police their own. Now put them in charge of it.

Why do you think your matched respondent chose to join the campaign to oppose defunding the police?



Matched Respondent's Donation Decision

We gave your matched respondent the opportunity to donate $10 to the **National Association for the Advancement of Colored People (NAACP)**, America's oldest and largest civil rights organization.

Below, we will ask you to guess whether or not your matched respondent donated $10 to the National Association for the Advancement of Colored People (NAACP).

Reminder: Your matched participant agreed to post the Tweet below on their Twitter account.

I have joined a campaign to oppose defunding the police: bit.ly/3DK3UEr. After joining, I was shown this article written by a Princeton professor on the strong scientific evidence that defunding the police would increase violent crime: https://www.washingtonpost.com/outlook/2020/06/12/defund-police-violent-crime/

washingtonpost.com

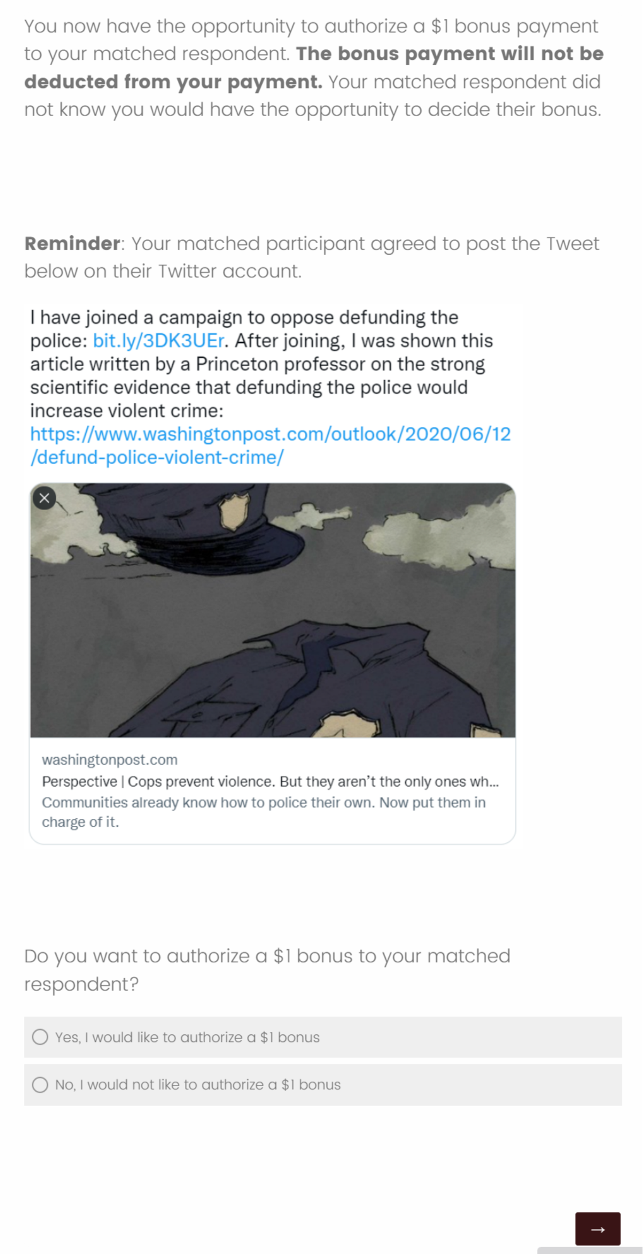
Perspective | Cops prevent violence. But they aren't the only ones wh...

Communities already know how to police their own. Now put them in charge of it.

Do you think that your matched participant chose to donate $5 to the National Association for the Advancement of Colored People (NAACP)?

○ Yes, I think my matched respondent chose to donate

○ No, I think my matched respondent **did not** choose to donate



You now have the opportunity to authorize a $1 bonus payment to your matched respondent. **The bonus payment will not be deducted from your payment.** Your matched respondent did not know you would have the opportunity to decide their bonus.

Reminder: Your matched participant agreed to post the Tweet below on their Twitter account.

I have joined a campaign to oppose defunding the police: bit.ly/3DK3UEr. After joining, I was shown this article written by a Princeton professor on the strong scientific evidence that defunding the police would increase violent crime: https://www.washingtonpost.com/outlook/2020/06/12/defund-police-violent-crime/

washingtonpost.com

Perspective | Cops prevent violence. But they aren't the only ones wh...

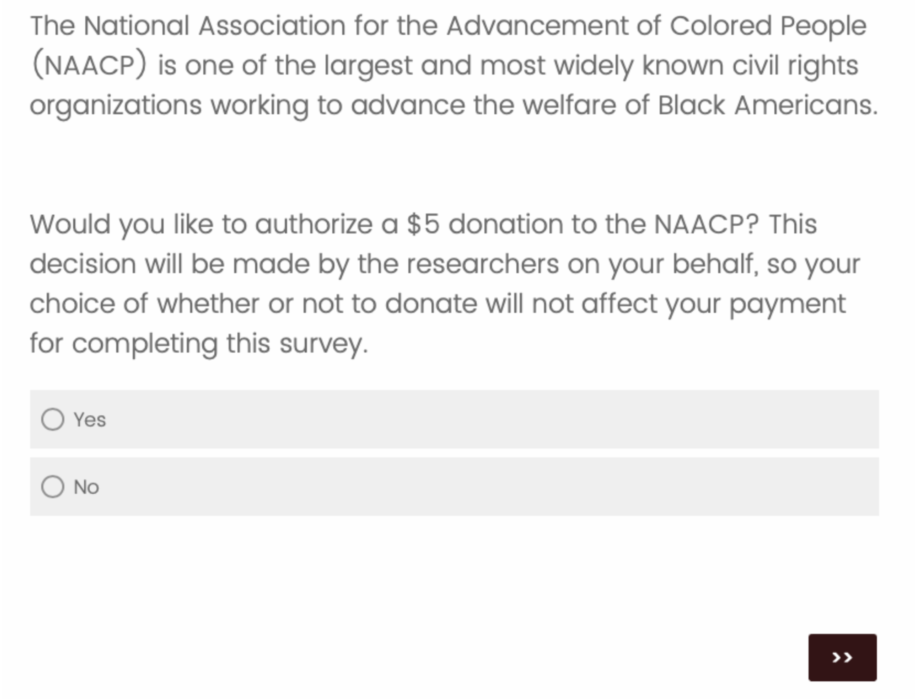
Communities already know how to police their own. Now put them in charge of it.

Do you want to authorize a $1 bonus to your matched respondent?

○ Yes, I would like to authorize a $1 bonus

○ No, I would not like to authorize a $1 bonus

E.2.4 Supporting experiment (willingness to donate to NAACP)



The National Association for the Advancement of Colored People (NAACP) is one of the largest and most widely known civil rights organizations working to advance the welfare of Black Americans.

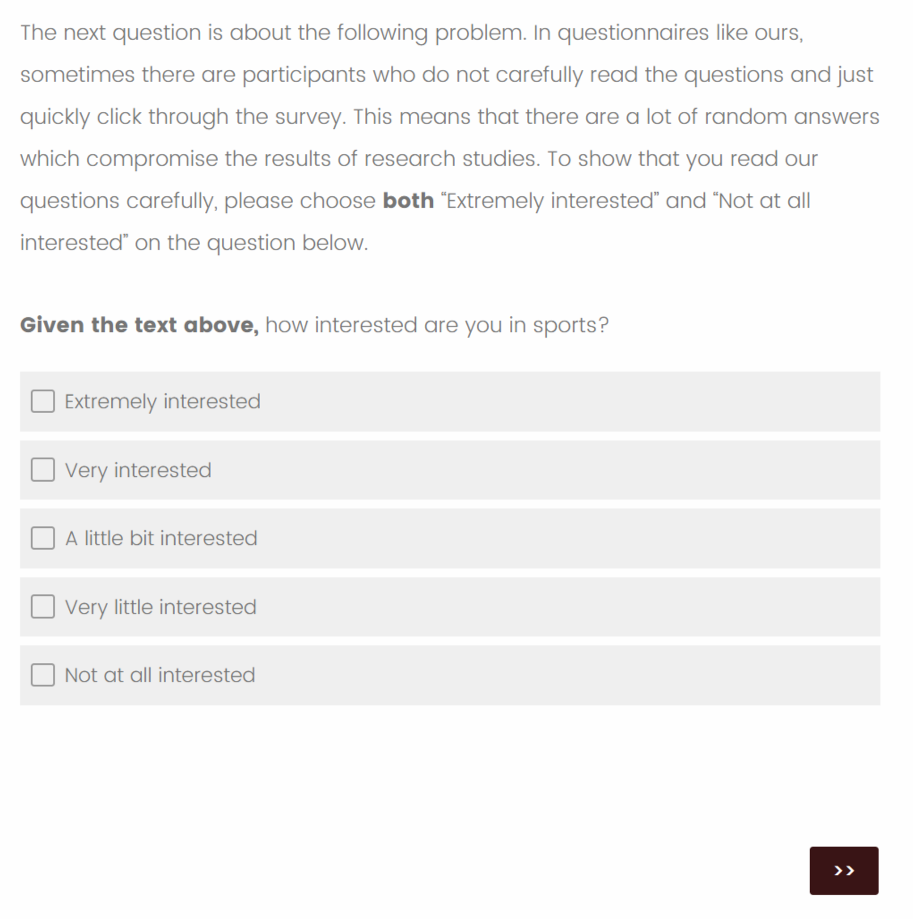
Would you like to authorize a $5 donation to the NAACP? This decision will be made by the researchers on your behalf, so your choice of whether or not to donate will not affect your payment for completing this survey.

○ Yes

○ No

E.3 Experiment 3: Expression of dissent – Republicans

E.3.1 Attention screener



The next question is about the following problem. In questionnaires like ours, sometimes there are participants who do not carefully read the questions and just quickly click through the survey. This means that there are a lot of random answers which compromise the results of research studies. To show that you read our questions carefully, please choose **both** "Extremely interested" and "Not at all interested" on the question below.

Given the text above, how interested are you in sports?

□ Extremely interested

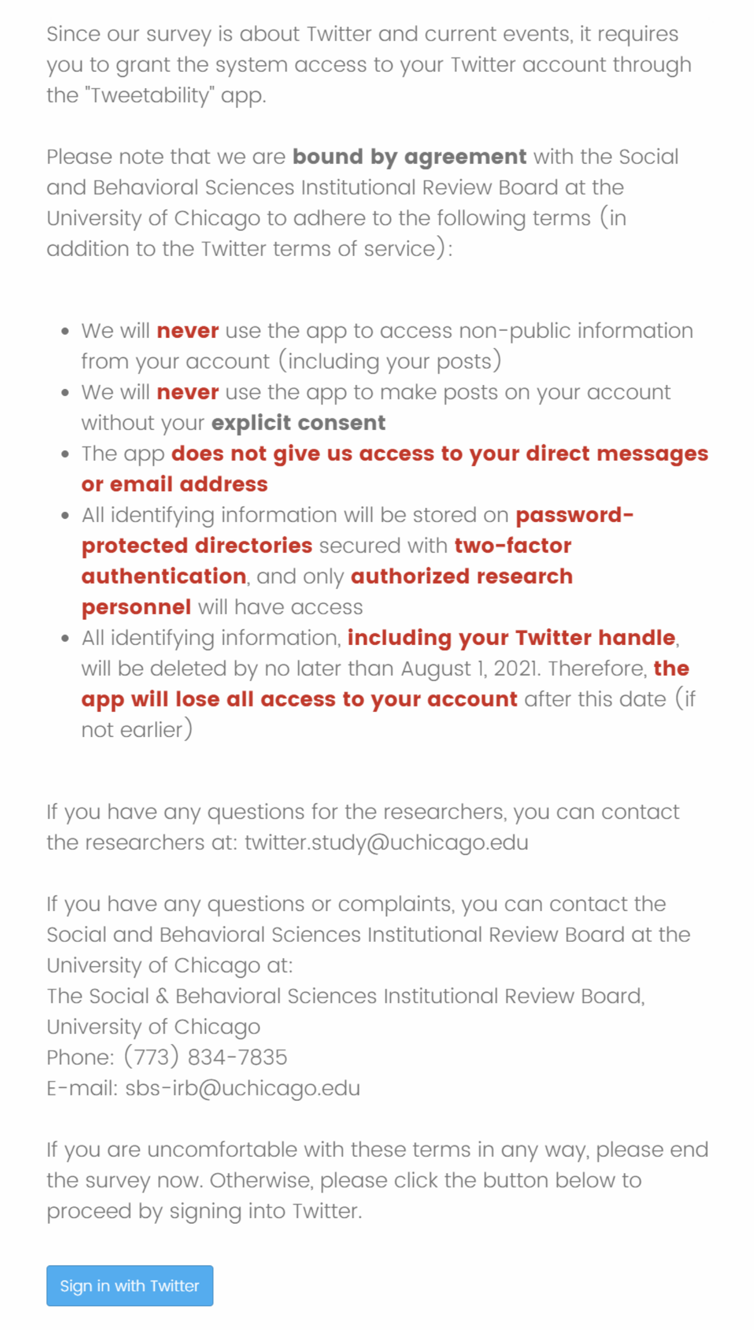
□ Very interested

□ A little bit interested

□ Very little interested

□ Not at all interested

E.3.2 Twitter information and login



Since our survey is about Twitter and current events, it requires you to grant the system access to your Twitter account through the "Tweetability" app.

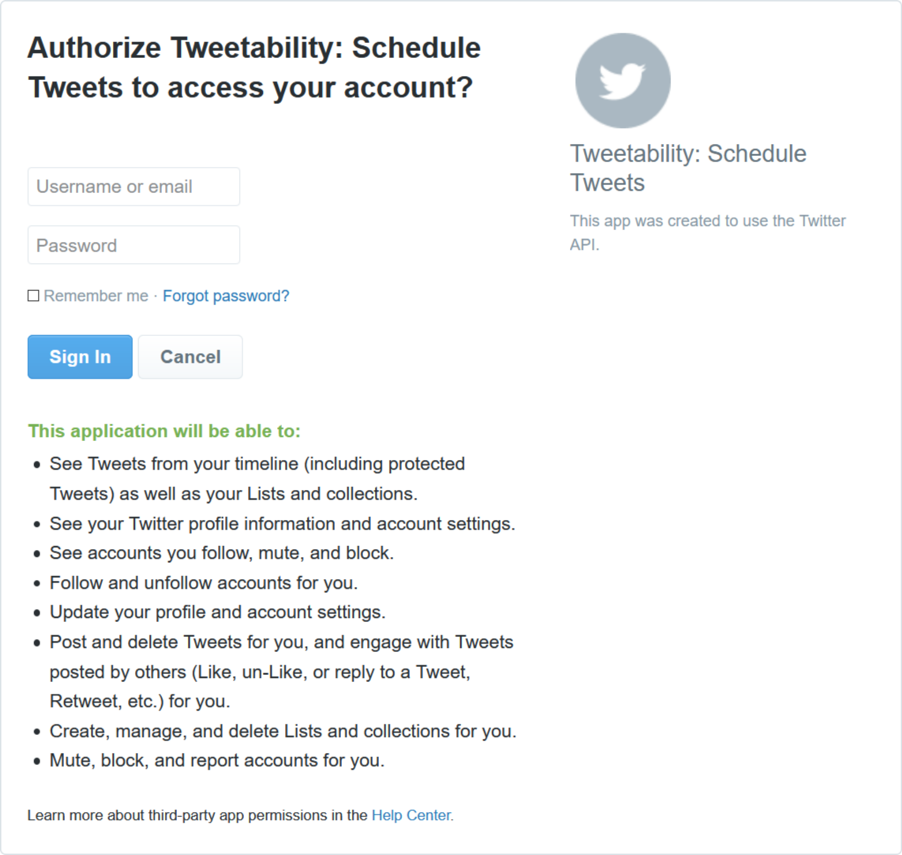
Please note that we are **bound by agreement** with the Social and Behavioral Sciences Institutional Review Board at the University of Chicago to adhere to the following terms (in addition to the Twitter terms of service):

- We will **never** use the app to access non-public information from your account (including your posts)
- We will **never** use the app to make posts on your account without your **explicit consent**
- The app **does not give us access to your direct messages or email address**
- All identifying information will be stored on **password-protected directories** secured with **two-factor authentication**, and only **authorized research personnel** will have access
- All identifying information, **including your Twitter handle**, will be deleted by no later than August 1, 2021. Therefore, **the app will lose all access to your account** after this date (if not earlier)

If you have any questions for the researchers, you can contact the researchers at: twitter.study@uchicago.edu

If you have any questions or complaints, you can contact the Social and Behavioral Sciences Institutional Review Board at the University of Chicago at:
The Social & Behavioral Sciences Institutional Review Board,
University of Chicago
Phone: (773) 834-7835
E-mail: sbs-irb@uchicago.edu

If you are uncomfortable with these terms in any way, please end the survey now. Otherwise, please click the button below to proceed by signing into Twitter.



Authorize Tweetability: Schedule Tweets to access your account?

Tweetability: Schedule Tweets

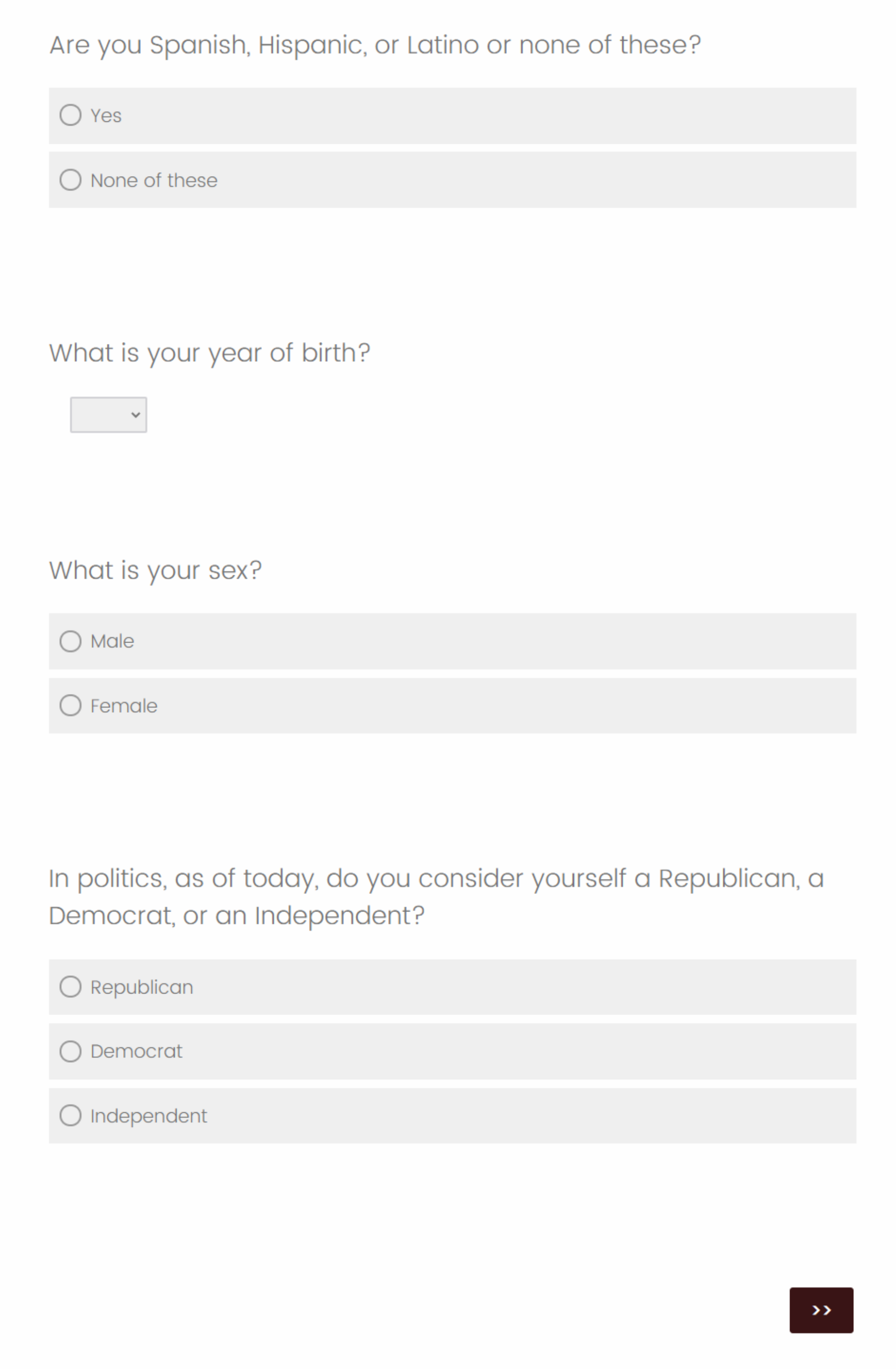
This app was created to use the Twitter API.

This application will be able to:

- See Tweets from your timeline (including protected Tweets) as well as your Lists and collections.
- See your Twitter profile information and account settings.
- See accounts you follow, mute, and block.
- Follow and unfollow accounts for you.
- Update your profile and account settings.
- Post and delete Tweets for you, and engage with Tweets posted by others (Like, un-Like, or reply to a Tweet, Retweet, etc.) for you.
- Create, manage, and delete Lists and collections for you.
- Mute, block, and report accounts for you.

Learn more about third-party app permissions in the Help Center.

E.3.3 Demographics



Are you Spanish, Hispanic, or Latino or none of these?

○ Yes

○ None of these

What is your year of birth?

What is your sex?

○ Male

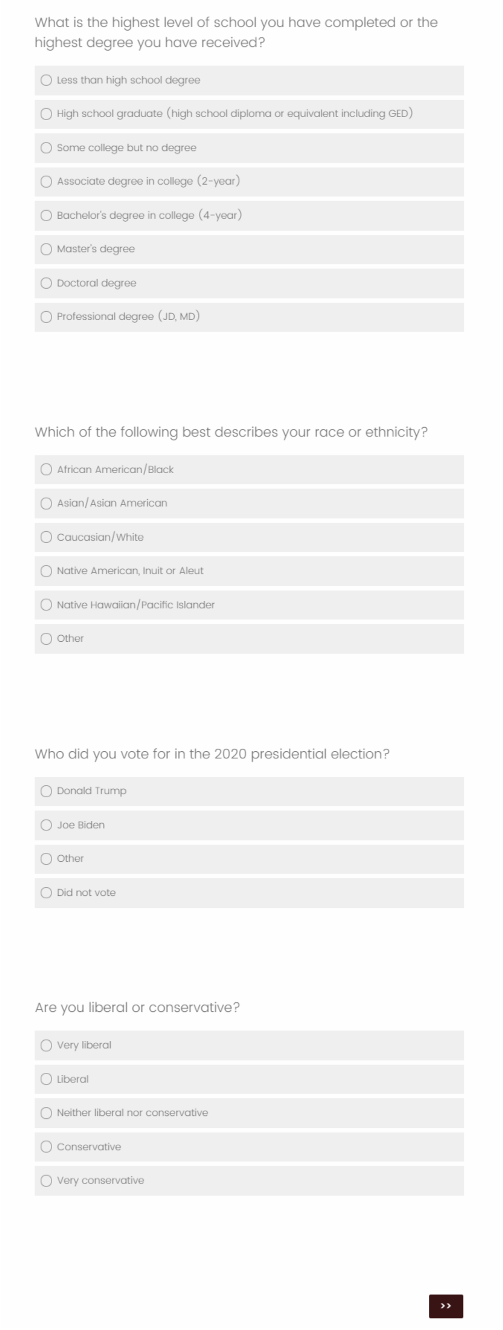
○ Female

In politics, as of today, do you consider yourself a Republican, a Democrat, or an Independent?

○ Republican

○ Democrat

○ Independent



What is the highest level of school you have completed or the highest degree you have received?

○ Less than high school degree

○ High school graduate (high school diploma or equivalent including GED)

○ Some college but no degree

○ Associate degree in college (2-year)

○ Bachelor's degree in college (4-year)

○ Master's degree

○ Doctoral degree

○ Professional degree (JD, MD)

Which of the following best describes your race or ethnicity?

○ African American/Black

○ Asian/Asian American

○ Caucasian/White

○ Native American, Inuit or Aleut

○ Native Hawaiian/Pacific Islander

○ Other

Who did you vote for in the 2020 presidential election?

○ Donald Trump

○ Joe Biden

○ Other

○ Did not vote

Are you liberal or conservative?

○ Very liberal

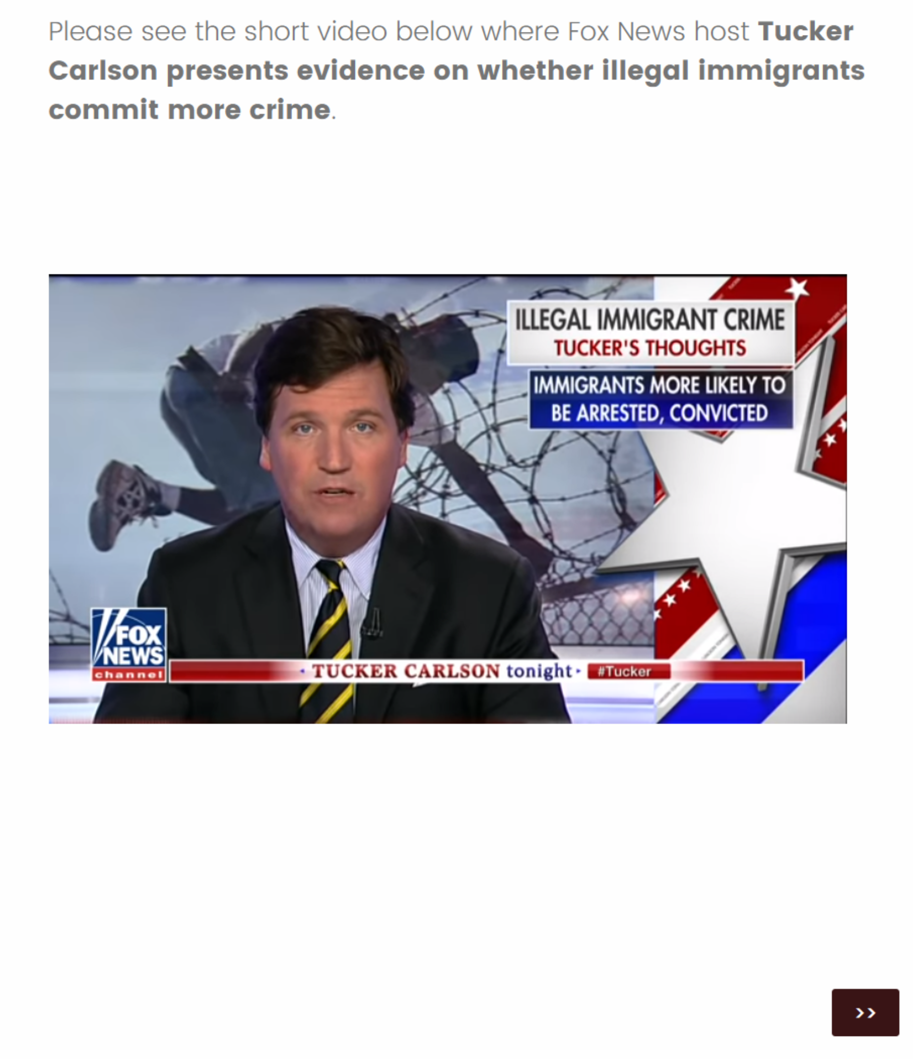
○ Liberal

○ Neither liberal nor conservative

○ Conservative

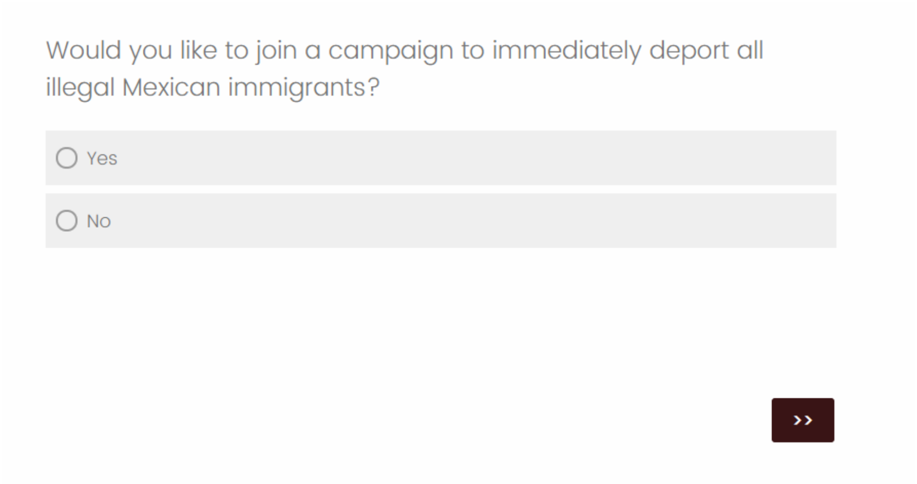
○ Very conservative

E.3.4 Video clip



Please see the short video below where Fox News host **Tucker Carlson presents evidence on whether illegal immigrants commit more crime**.

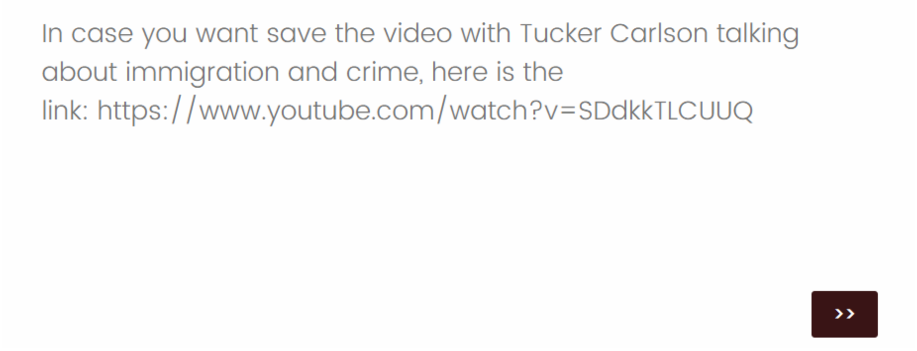
E.3.5 Treatment: “After” wording



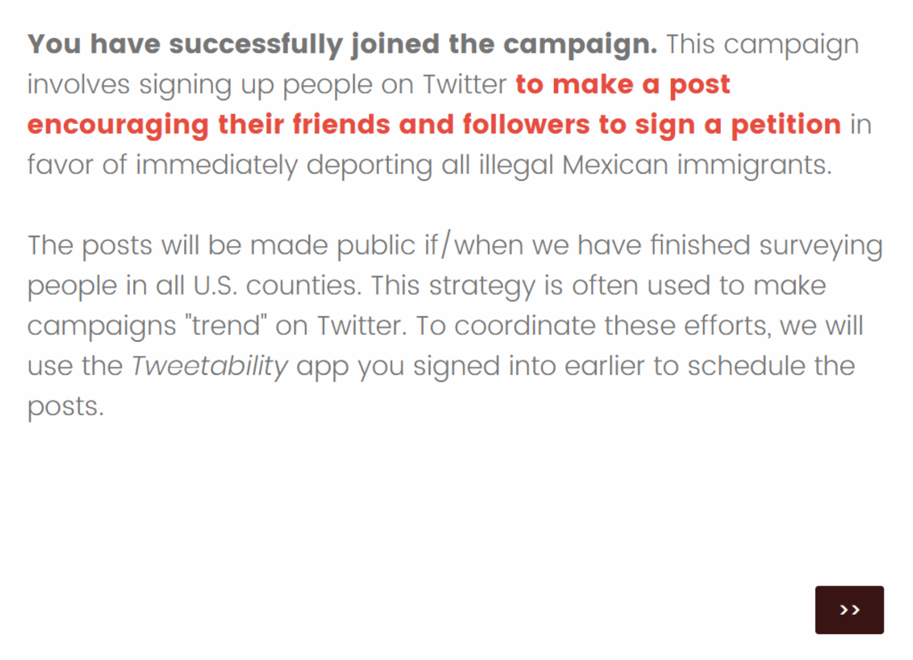
Would you like to join a campaign to immediately deport all illegal Mexican immigrants?

○ Yes

○ No

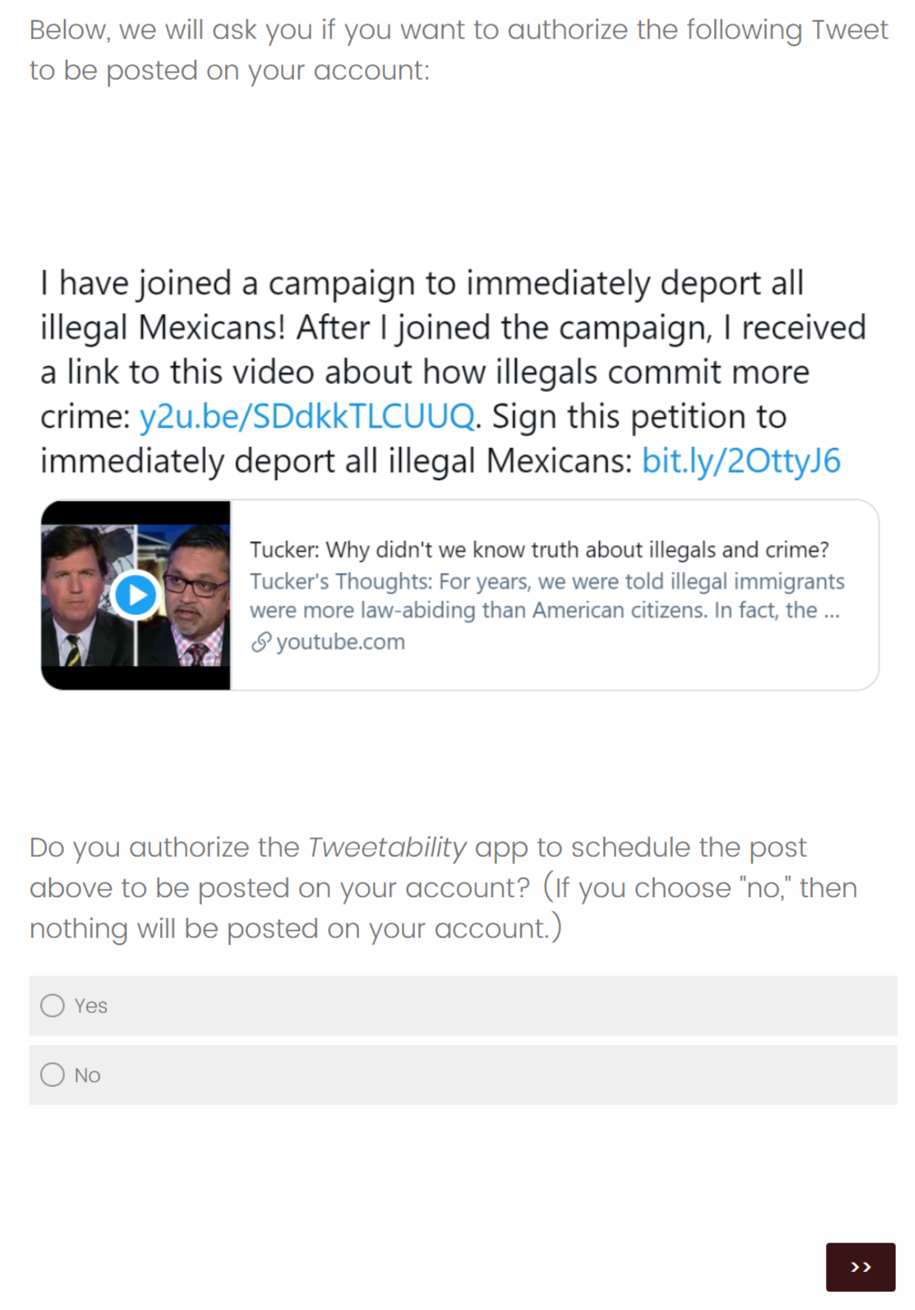


In case you want save the video with Tucker Carlson talking about immigration and crime, here is the link: https://www.youtube.com/watch?v=SDdkkTLCUUQ



You have successfully joined the campaign. This campaign involves signing up people on Twitter **to make a post encouraging their friends and followers to sign a petition** in favor of immediately deporting all illegal Mexican immigrants.

The posts will be made public if/when we have finished surveying people in all U.S. counties. This strategy is often used to make campaigns "trend" on Twitter. To coordinate these efforts, we will use the *Tweetability* app you signed into earlier to schedule the posts.



Below, we will ask you if you want to authorize the following Tweet to be posted on your account:

I have joined a campaign to immediately deport all illegal Mexicans! After I joined the campaign, I received a link to this video about how illegals commit more crime: y2u.be/SDdkkTLCUUQ. Sign this petition to immediately deport all illegal Mexicans: bit.ly/2OttyJ6

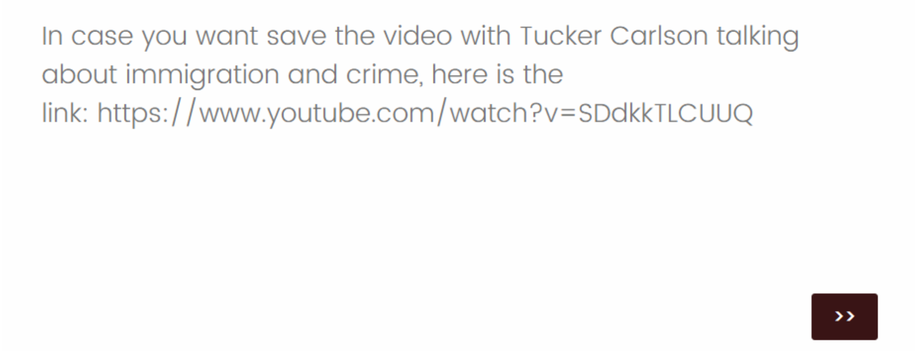
Tucker: Why didn't we know truth about illegals and crime?
Tucker's Thoughts: For years, we were told illegal immigrants were more law-abiding than American citizens. In fact, the ...
youtube.com

Do you authorize the *Tweetability* app to schedule the post above to be posted on your account? (If you choose "no," then nothing will be posted on your account.)

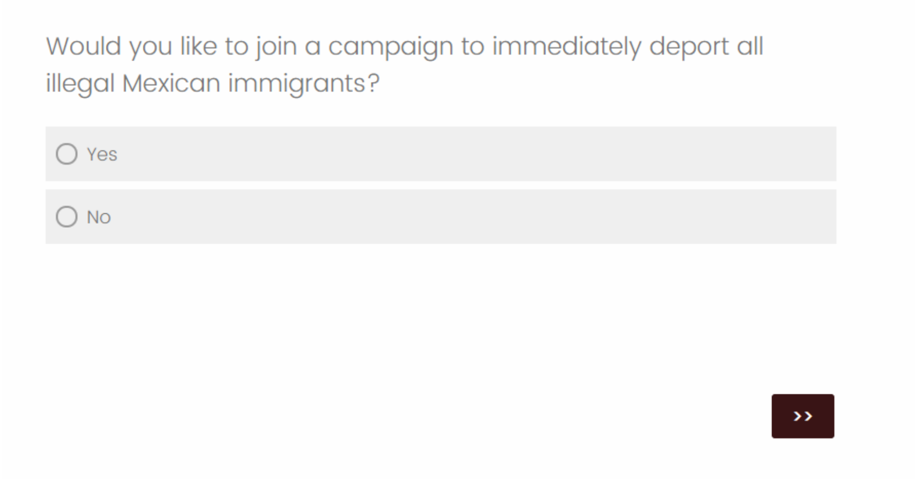
○ Yes

○ No

E.3.6 Treatment: “Before” wording



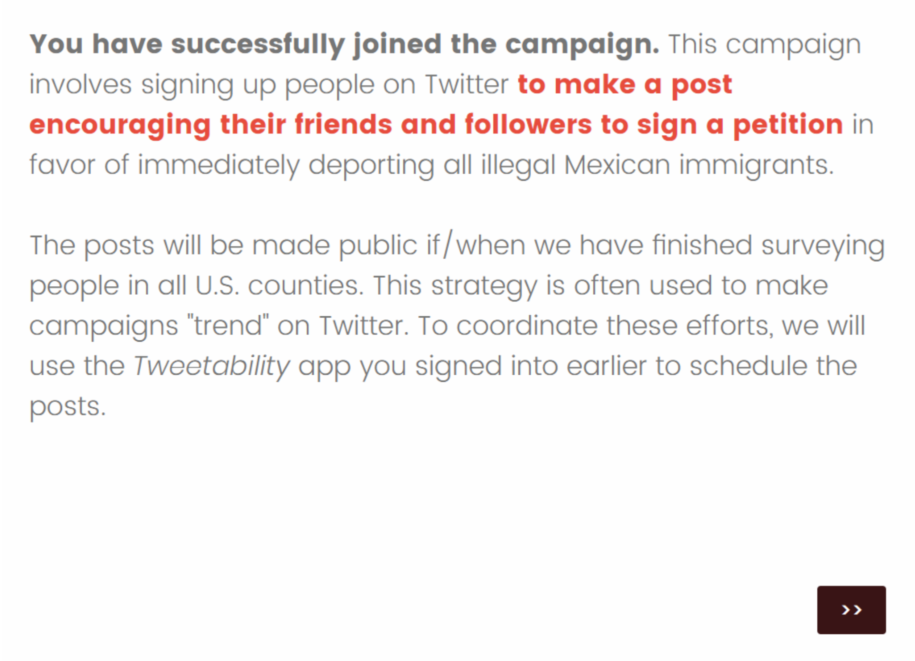
In case you want save the video with Tucker Carlson talking about immigration and crime, here is the link: https://www.youtube.com/watch?v=SDdkkTLCUUQ



Would you like to join a campaign to immediately deport all illegal Mexican immigrants?

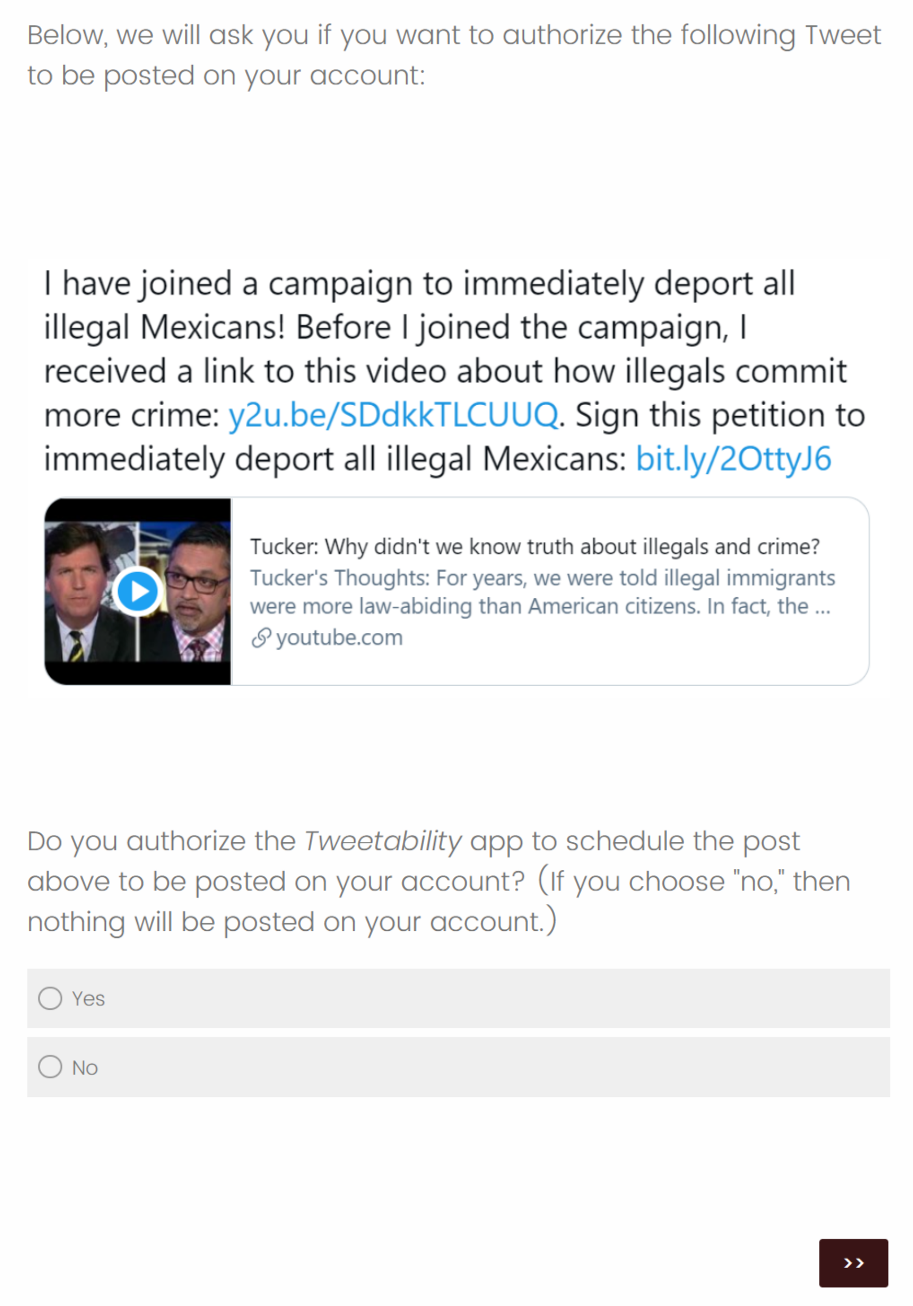
○ Yes

○ No



You have successfully joined the campaign. This campaign involves signing up people on Twitter **to make a post encouraging their friends and followers to sign a petition** in favor of immediately deporting all illegal Mexican immigrants.

The posts will be made public if/when we have finished surveying people in all U.S. counties. This strategy is often used to make campaigns "trend" on Twitter. To coordinate these efforts, we will use the *Tweetability* app you signed into earlier to schedule the posts.



Below, we will ask you if you want to authorize the following Tweet to be posted on your account:

I have joined a campaign to immediately deport all illegal Mexicans! Before I joined the campaign, I received a link to this video about how illegals commit more crime: y2u.be/SDdkkTLCUUQ. Sign this petition to immediately deport all illegal Mexicans: bit.ly/2OttyJ6

Tucker: Why didn't we know truth about illegals and crime?
Tucker's Thoughts: For years, we were told illegal immigrants were more law-abiding than American citizens. In fact, the ...
🔗 youtube.com

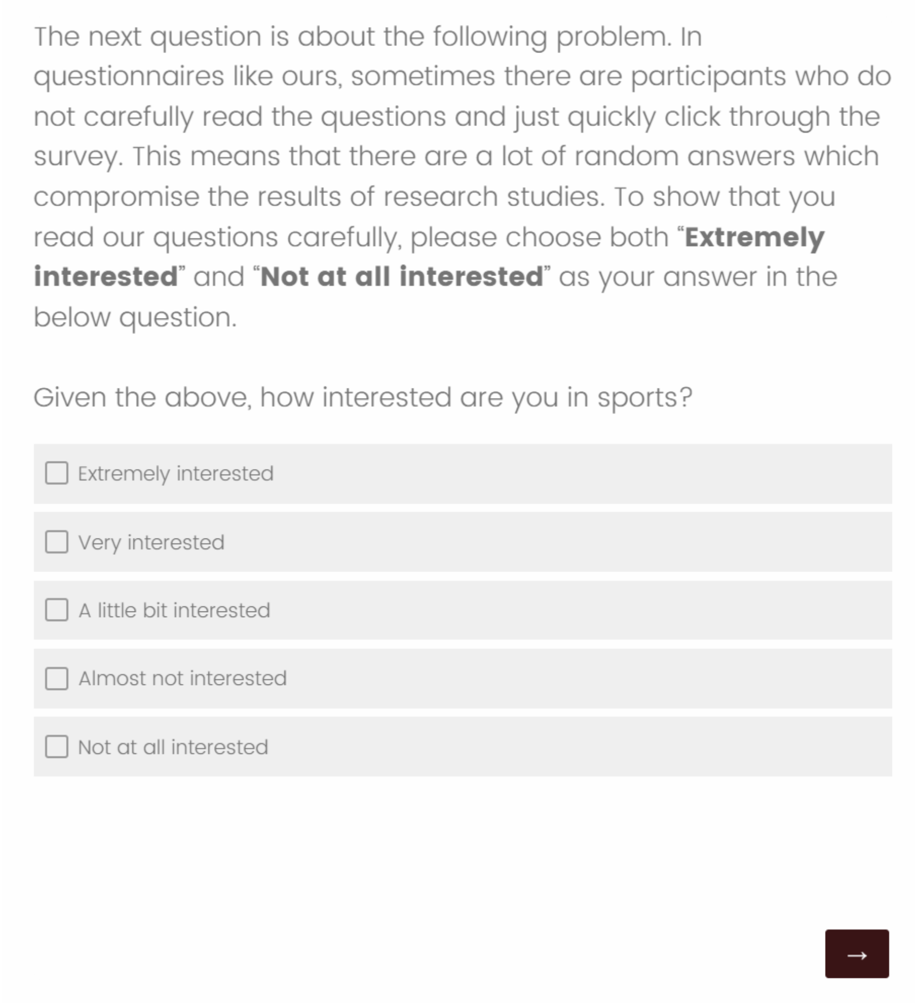
Do you authorize the *Tweetability* app to schedule the post above to be posted on your account? (If you choose "no," then nothing will be posted on your account.)

○ Yes

○ No

E.4 Experiment 4: Interpretation of dissent – Democrats

E.4.1 Attention screener and background questions



The next question is about the following problem. In questionnaires like ours, sometimes there are participants who do not carefully read the questions and just quickly click through the survey. This means that there are a lot of random answers which compromise the results of research studies. To show that you read our questions carefully, please choose both "**Extremely interested**" and "**Not at all interested**" as your answer in the below question.

Given the above, how interested are you in sports?

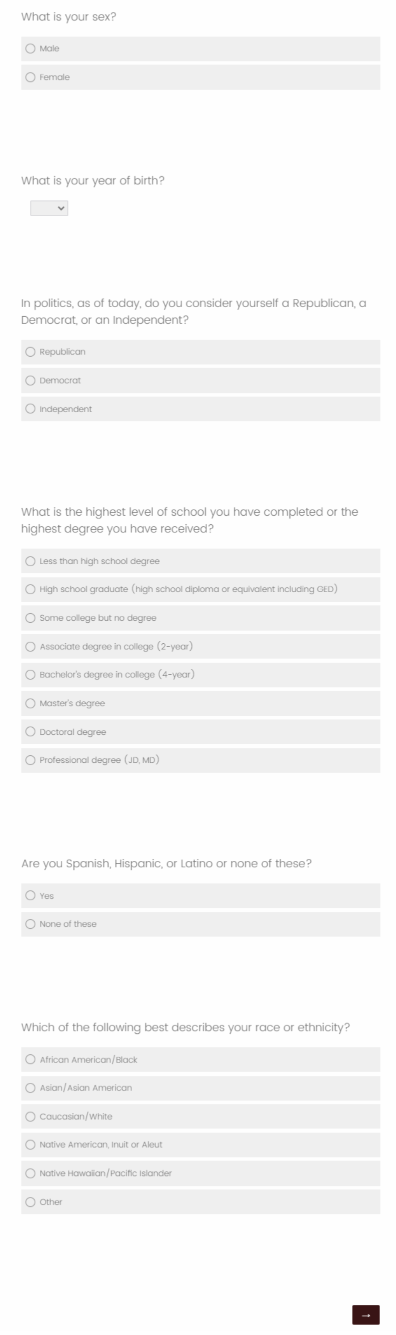
□ Extremely interested

□ Very interested

□ A little bit interested

□ Almost not interested

□ Not at all interested



What is your sex?

○ Male

○ Female

What is your year of birth?

In politics, as of today, do you consider yourself a Republican, a Democrat, or an Independent?

○ Republican

○ Democrat

○ Independent

What is the highest level of school you have completed or the highest degree you have received?

○ Less than high school degree

○ High school graduate (high school diploma or equivalent including GED)

○ Some college but no degree

○ Associate degree in college (2-year)

○ Bachelor's degree in college (4-year)

○ Master's degree

○ Doctoral degree

○ Professional degree (JD, MD)

Are you Spanish, Hispanic, or Latino or none of these?

○ Yes

○ None of these

Which of the following best describes your race or ethnicity?

○ African American/Black

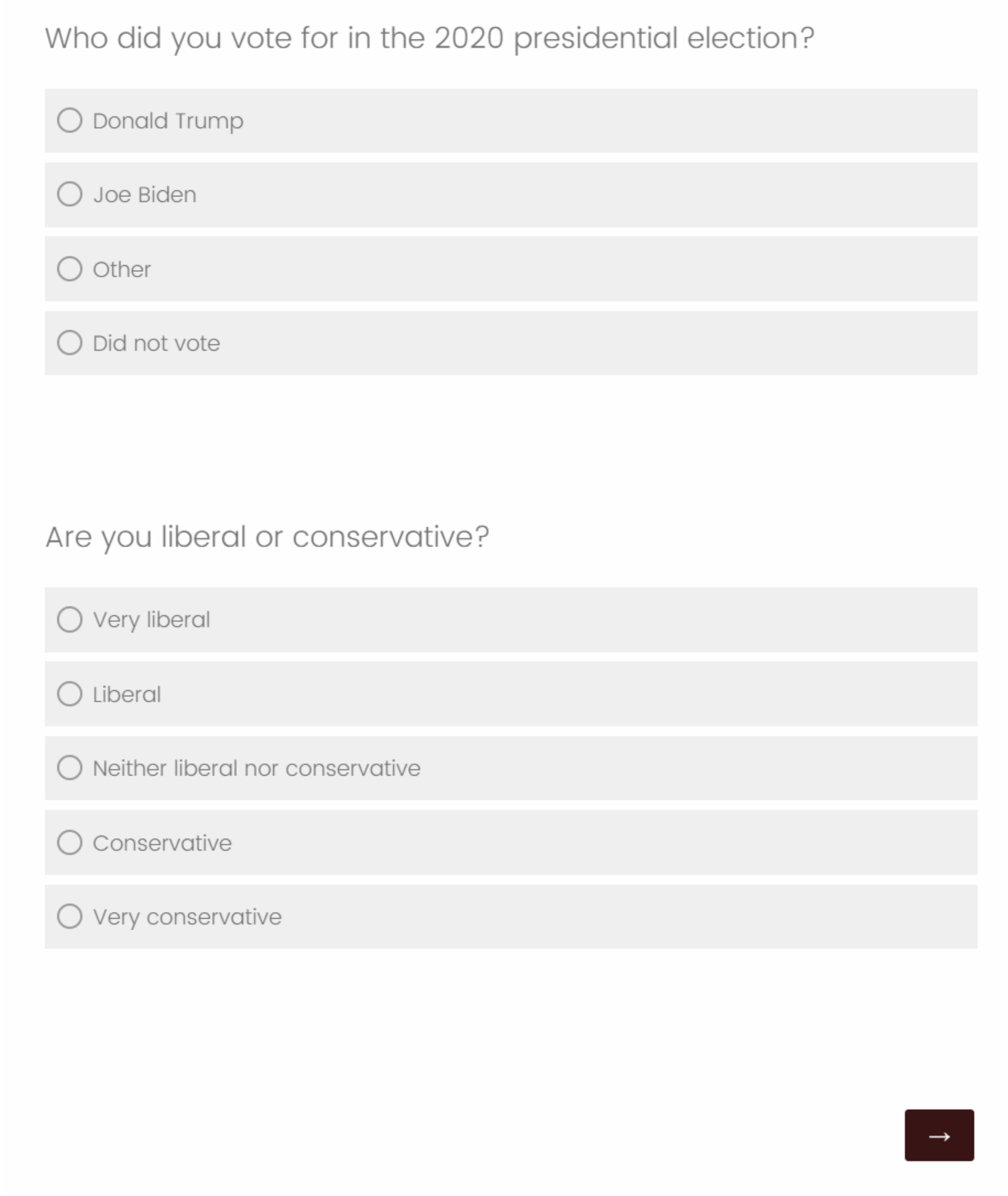
○ Asian/Asian American

○ Caucasian/White

○ Native American, Inuit or Aleut

○ Native Hawaiian/Pacific Islander

○ Other



Who did you vote for in the 2020 presidential election?

○ Donald Trump

○ Joe Biden

○ Other

○ Did not vote

Are you liberal or conservative?

○ Very liberal

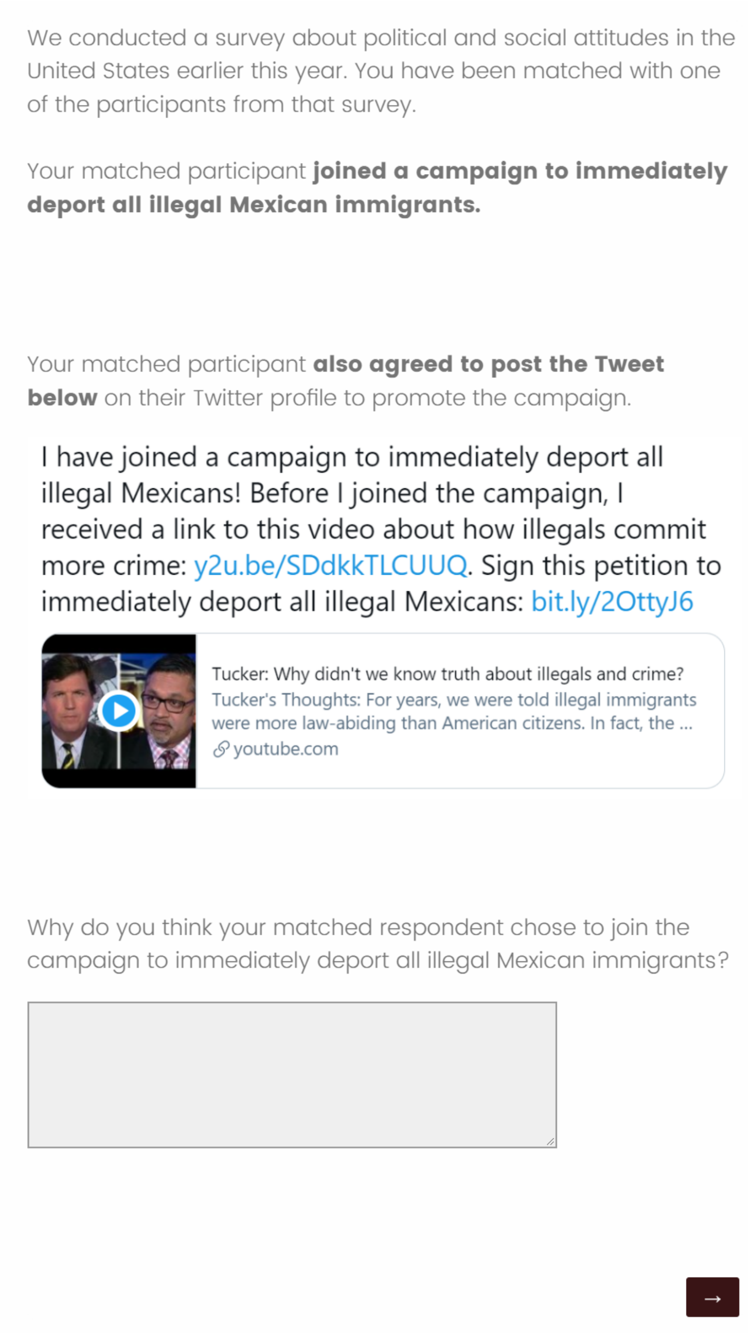
○ Liberal

○ Neither liberal nor conservative

○ Conservative

○ Very conservative

E.4.2 Treatment: “Before” condition (rationale)



We conducted a survey about political and social attitudes in the United States earlier this year. You have been matched with one of the participants from that survey.

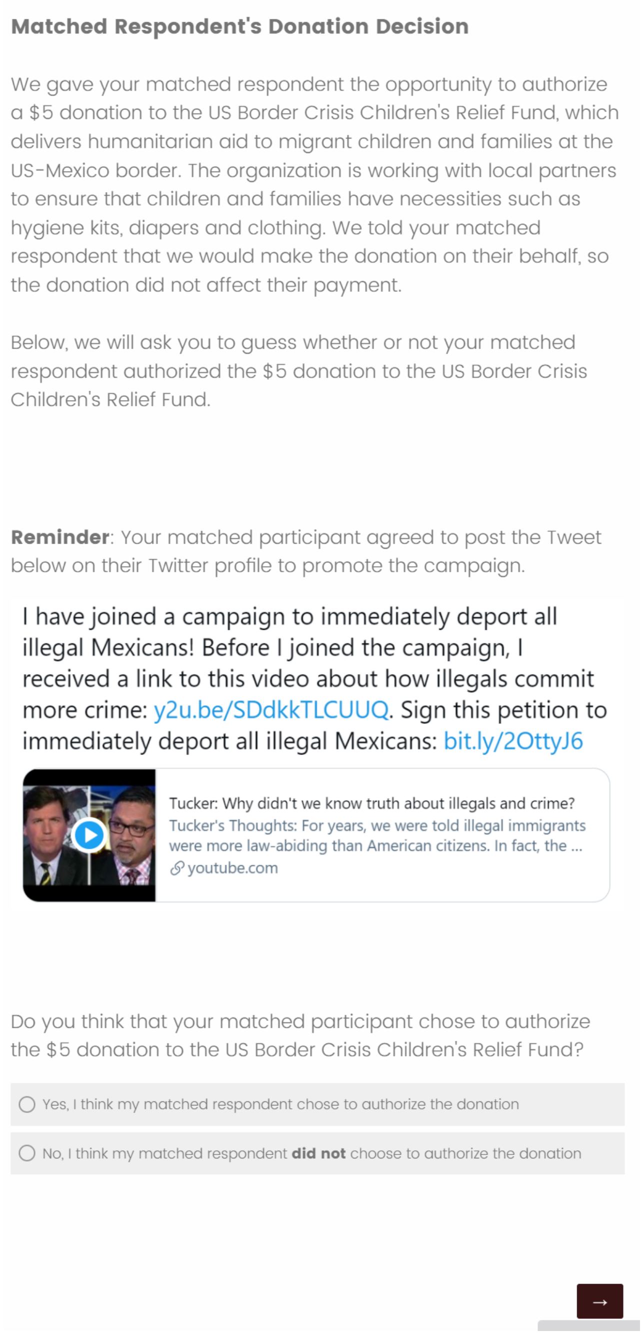
Your matched participant **joined a campaign to immediately deport all illegal Mexican immigrants.**

Your matched participant **also agreed to post the Tweet below** on their Twitter profile to promote the campaign.

I have joined a campaign to immediately deport all illegal Mexicans! Before I joined the campaign, I received a link to this video about how illegals commit more crime: y2u.be/SDdkkTLCUUQ. Sign this petition to immediately deport all illegal Mexicans: bit.ly/2OttyJ6

Tucker: Why didn't we know truth about illegals and crime?
Tucker's Thoughts: For years, we were told illegal immigrants were more law-abiding than American citizens. In fact, the ...
youtube.com

Why do you think your matched respondent chose to join the campaign to immediately deport all illegal Mexican immigrants?



Matched Respondent's Donation Decision

We gave your matched respondent the opportunity to authorize a $5 donation to the US Border Crisis Children's Relief Fund, which delivers humanitarian aid to migrant children and families at the US-Mexico border. The organization is working with local partners to ensure that children and families have necessities such as hygiene kits, diapers and clothing. We told your matched respondent that we would make the donation on their behalf, so the donation did not affect their payment.

Below, we will ask you to guess whether or not your matched respondent authorized the $5 donation to the US Border Crisis Children's Relief Fund.

Reminder: Your matched participant agreed to post the Tweet below on their Twitter profile to promote the campaign.

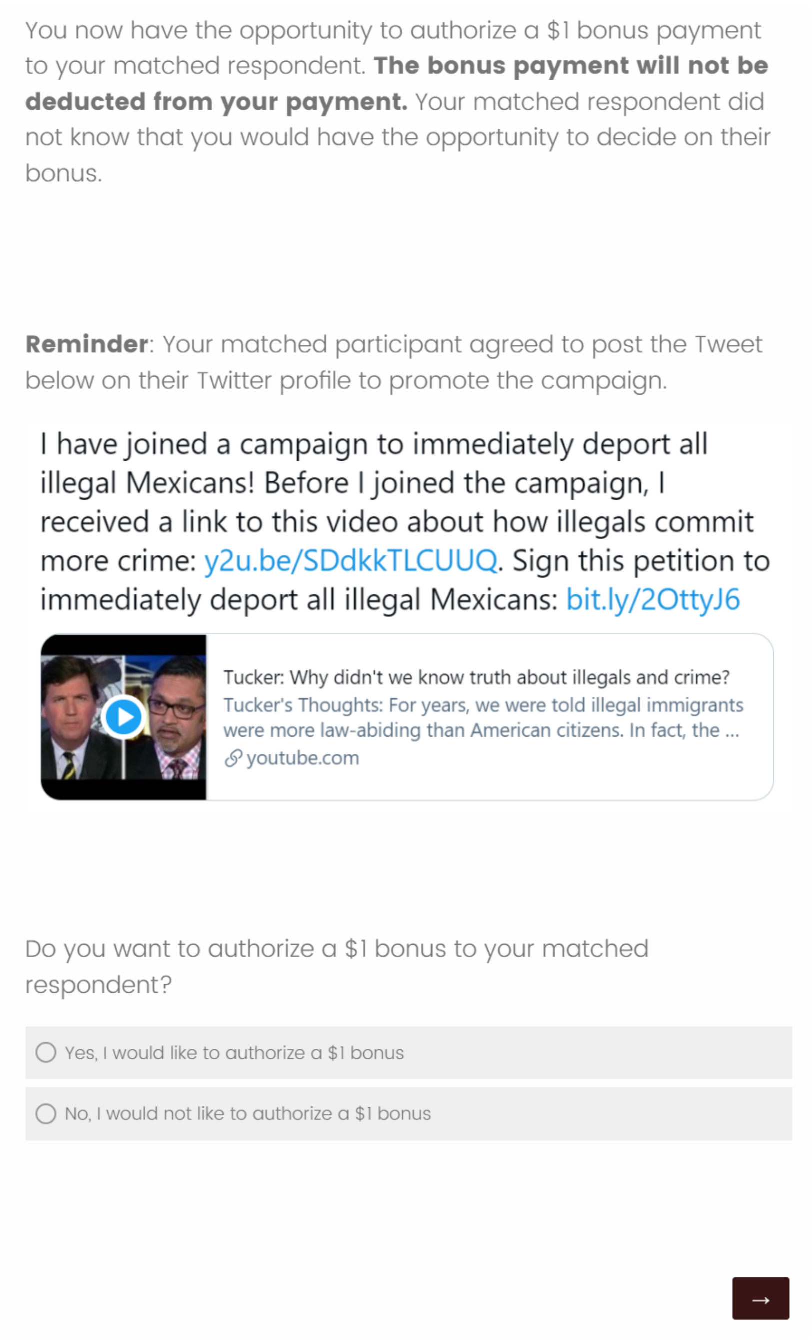
I have joined a campaign to immediately deport all illegal Mexicans! Before I joined the campaign, I received a link to this video about how illegals commit more crime: y2u.be/SDdkkTLCUUQ. Sign this petition to immediately deport all illegal Mexicans: bit.ly/2OttyJ6

Tucker: Why didn't we know truth about illegals and crime?
Tucker's Thoughts: For years, we were told illegal immigrants were more law-abiding than American citizens. In fact, the ...
youtube.com

Do you think that your matched participant chose to authorize the $5 donation to the US Border Crisis Children's Relief Fund?

○ Yes, I think my matched respondent chose to authorize the donation

○ No, I think my matched respondent **did not** choose to authorize the donation



You now have the opportunity to authorize a $1 bonus payment to your matched respondent. **The bonus payment will not be deducted from your payment.** Your matched respondent did not know that you would have the opportunity to decide on their bonus.

Reminder: Your matched participant agreed to post the Tweet below on their Twitter profile to promote the campaign.

I have joined a campaign to immediately deport all illegal Mexicans! Before I joined the campaign, I received a link to this video about how illegals commit more crime: y2u.be/SDdkkTLCUUQ. Sign this petition to immediately deport all illegal Mexicans: bit.ly/2OttyJ6

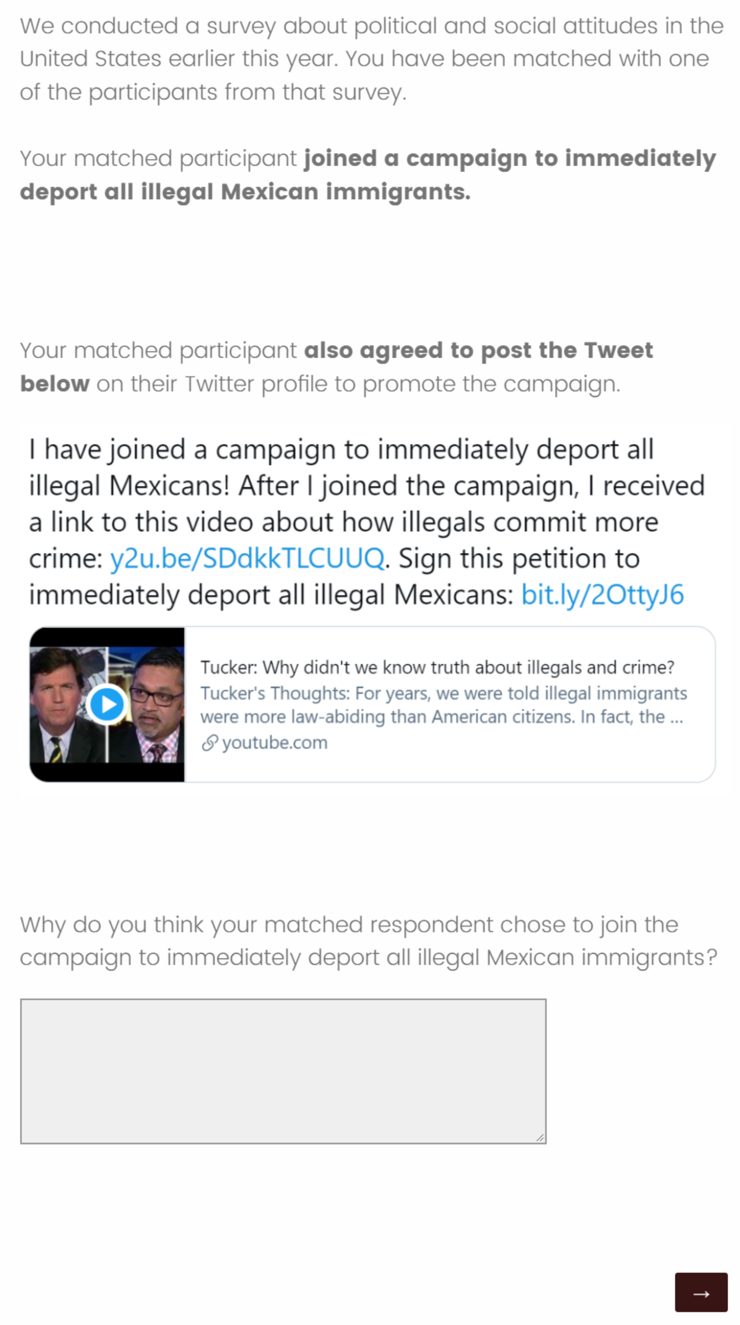
Tucker: Why didn't we know truth about illegals and crime?
Tucker's Thoughts: For years, we were told illegal immigrants were more law-abiding than American citizens. In fact, the ...
youtube.com

Do you want to authorize a $1 bonus to your matched respondent?

○ Yes, I would like to authorize a $1 bonus

○ No, I would not like to authorize a $1 bonus

E.4.3 Treatment: "After" condition (no rationale)



We conducted a survey about political and social attitudes in the United States earlier this year. You have been matched with one of the participants from that survey.

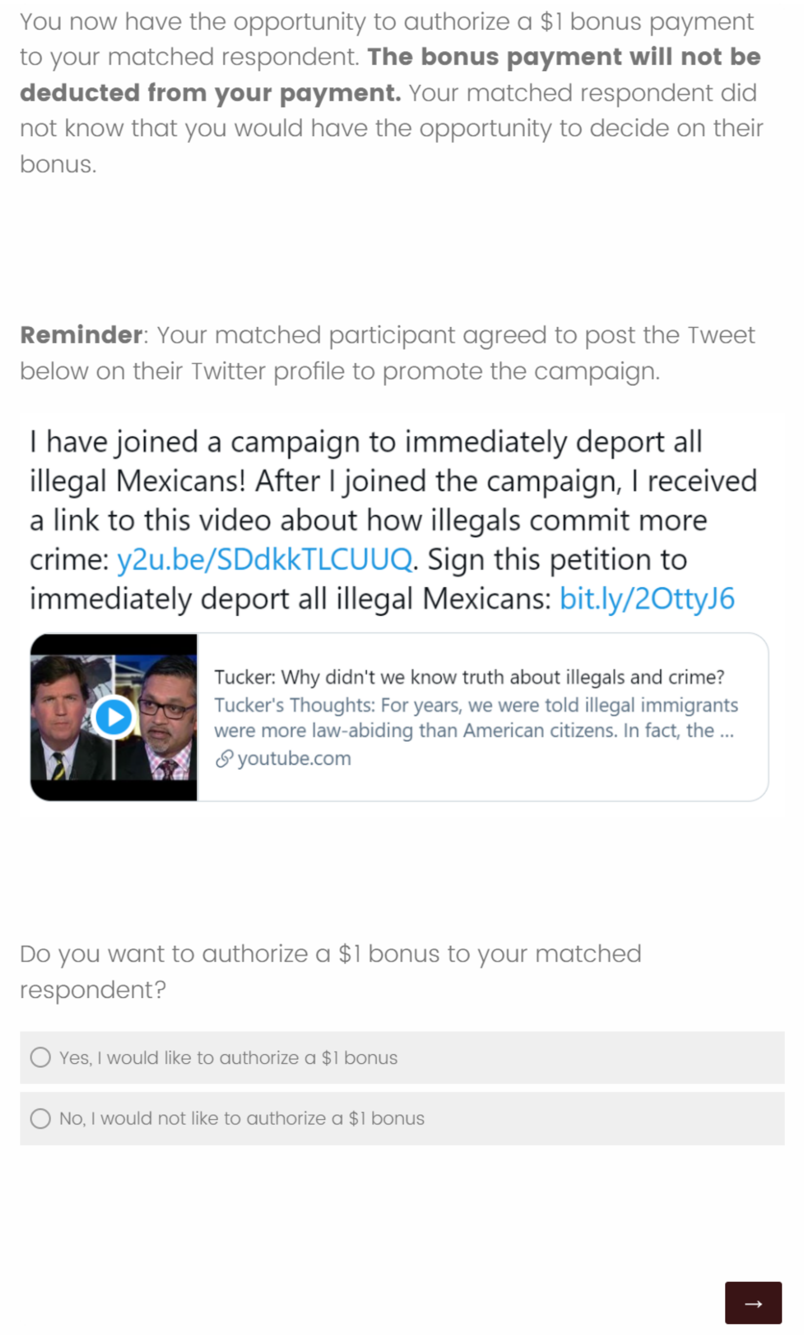
Your matched participant **joined a campaign to immediately deport all illegal Mexican immigrants.**

Your matched participant **also agreed to post the Tweet below** on their Twitter profile to promote the campaign.

I have joined a campaign to immediately deport all illegal Mexicans! After I joined the campaign, I received a link to this video about how illegals commit more crime: y2u.be/SDdkkTLCUUQ. Sign this petition to immediately deport all illegal Mexicans: bit.ly/2OttyJ6

Tucker: Why didn't we know truth about illegals and crime?
Tucker's Thoughts: For years, we were told illegal immigrants were more law-abiding than American citizens. In fact, the ...
youtube.com

Why do you think your matched respondent chose to join the campaign to immediately deport all illegal Mexican immigrants?



You now have the opportunity to authorize a $1 bonus payment to your matched respondent. **The bonus payment will not be deducted from your payment.** Your matched respondent did not know that you would have the opportunity to decide on their bonus.

Reminder: Your matched participant agreed to post the Tweet below on their Twitter profile to promote the campaign.

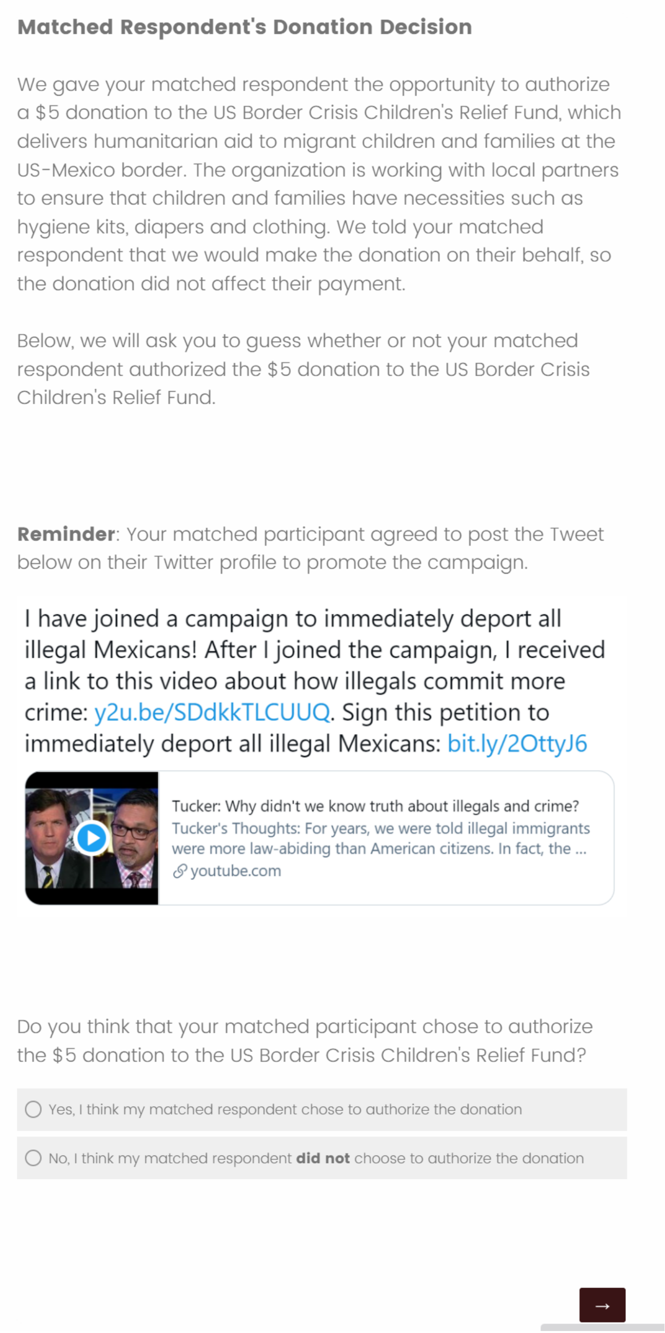
I have joined a campaign to immediately deport all illegal Mexicans! After I joined the campaign, I received a link to this video about how illegals commit more crime: y2u.be/SDdkkTLCUUQ. Sign this petition to immediately deport all illegal Mexicans: bit.ly/2OttyJ6

Tucker: Why didn't we know truth about illegals and crime?
Tucker's Thoughts: For years, we were told illegal immigrants were more law-abiding than American citizens. In fact, the ...
youtube.com

Do you want to authorize a $1 bonus to your matched respondent?

○ Yes, I would like to authorize a $1 bonus

○ No, I would not like to authorize a $1 bonus



Matched Respondent's Donation Decision

We gave your matched respondent the opportunity to authorize a $5 donation to the US Border Crisis Children's Relief Fund, which delivers humanitarian aid to migrant children and families at the US-Mexico border. The organization is working with local partners to ensure that children and families have necessities such as hygiene kits, diapers and clothing. We told your matched respondent that we would make the donation on their behalf, so the donation did not affect their payment.

Below, we will ask you to guess whether or not your matched respondent authorized the $5 donation to the US Border Crisis Children's Relief Fund.

Reminder: Your matched participant agreed to post the Tweet below on their Twitter profile to promote the campaign.

I have joined a campaign to immediately deport all illegal Mexicans! After I joined the campaign, I received a link to this video about how illegals commit more crime: y2u.be/SDdkkTLCUUQ. Sign this petition to immediately deport all illegal Mexicans: bit.ly/2OttyJ6

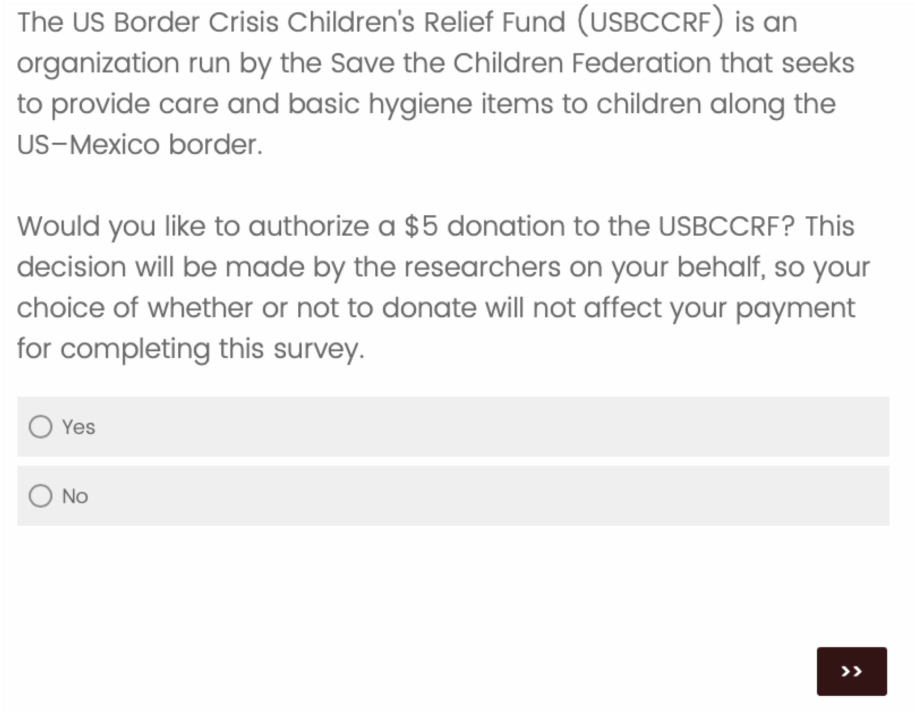
Tucker: Why didn't we know truth about illegals and crime?
Tucker's Thoughts: For years, we were told illegal immigrants were more law-abiding than American citizens. In fact, the ...
youtube.com

Do you think that your matched participant chose to authorize the $5 donation to the US Border Crisis Children's Relief Fund?

○ Yes, I think my matched respondent chose to authorize the donation

○ No, I think my matched respondent **did not** choose to authorize the donation

E.4.4 Supporting experiment (willingness to donate to USBCCRF)



The US Border Crisis Children's Relief Fund (USBCCRF) is an organization run by the Save the Children Federation that seeks to provide care and basic hygiene items to children along the US–Mexico border.

Would you like to authorize a $5 donation to the USBCCRF? This decision will be made by the researchers on your behalf, so your choice of whether or not to donate will not affect your payment for completing this survey.

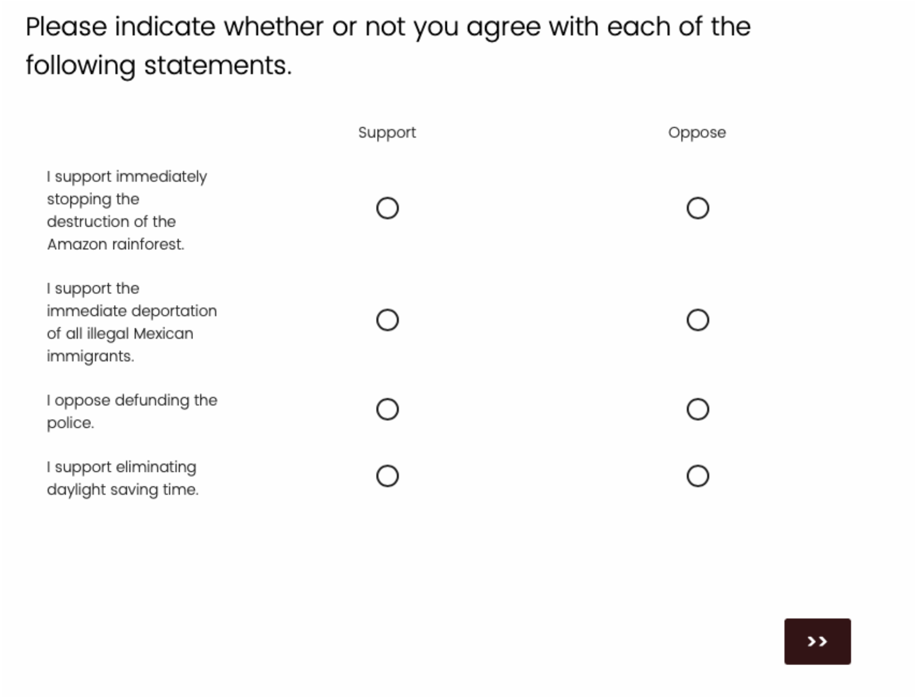
○ Yes

○ No

E.5 Survey: Anticipated social sanctions

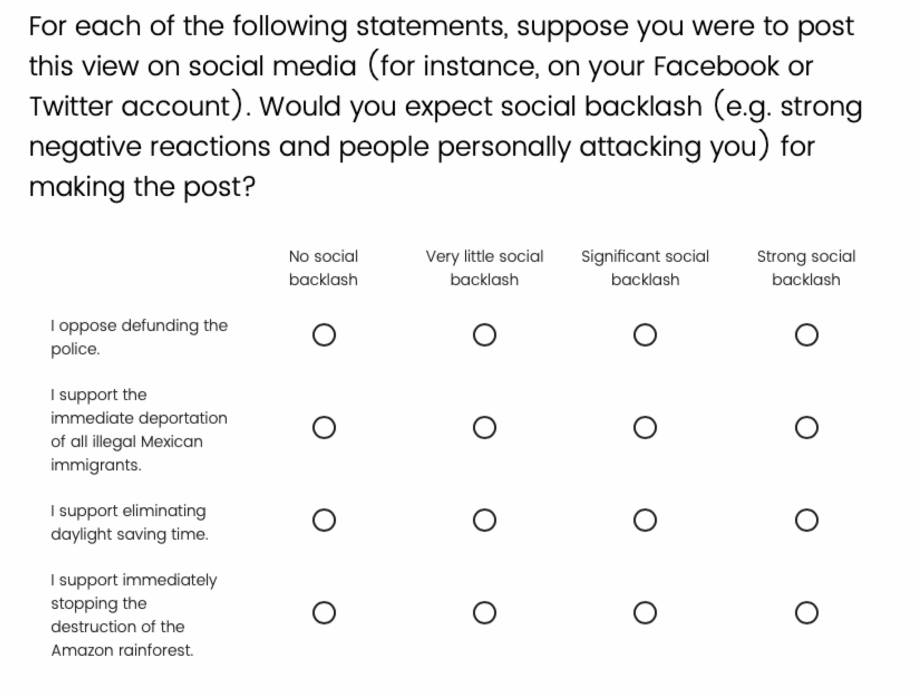
Note: This survey included the same demographic questions as the main experiments. We do not repeat screenshots of the demographic questions for the auxiliary survey and experiments.

E.5.1 Support and Anticipated Sanctions



Please indicate whether or not you agree with each of the following statements.

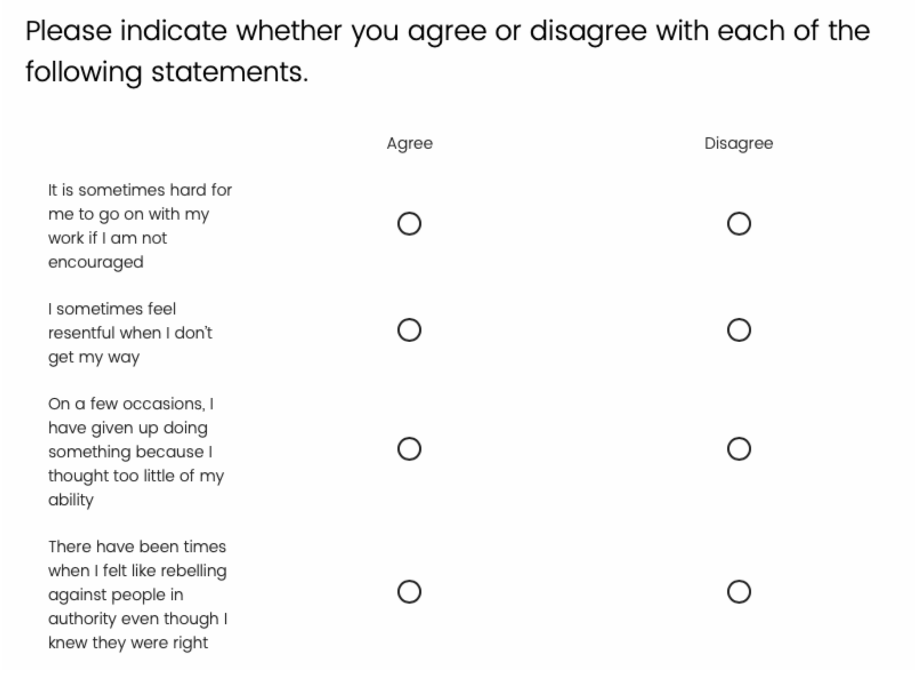
	Support	Oppose
I support immediately stopping the destruction of the Amazon rainforest.	○	○
I support the immediate deportation of all illegal Mexican immigrants.	○	○
I oppose defunding the police.	○	○
I support eliminating daylight saving time.	○	○



For each of the following statements, suppose you were to post this view on social media (for instance, on your Facebook or Twitter account). Would you expect social backlash (e.g. strong negative reactions and people personally attacking you) for making the post?

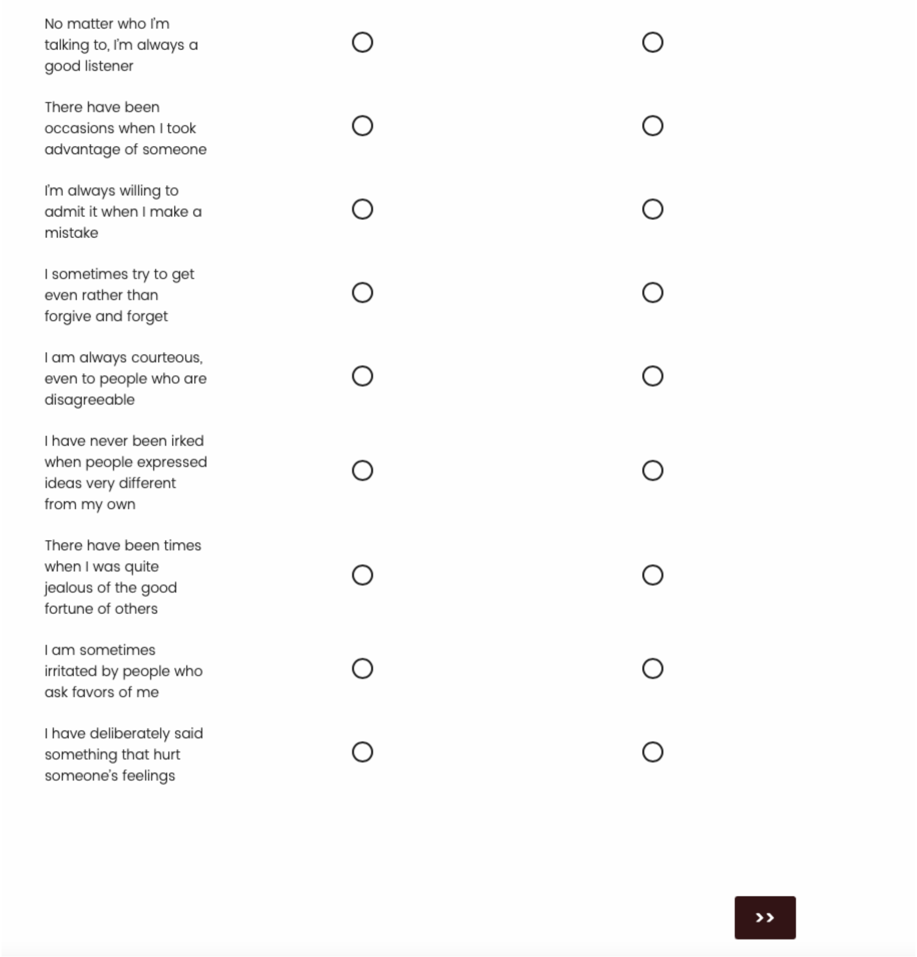
	No social backlash	Very little social backlash	Significant social backlash	Strong social backlash
I oppose defunding the police.	○	○	○	○
I support the immediate deportation of all illegal Mexican immigrants.	○	○	○	○
I support eliminating daylight saving time.	○	○	○	○
I support immediately stopping the destruction of the Amazon rainforest.	○	○	○	○

E.5.2 Marlowe-Crowne scale



Please indicate whether you agree or disagree with each of the following statements.

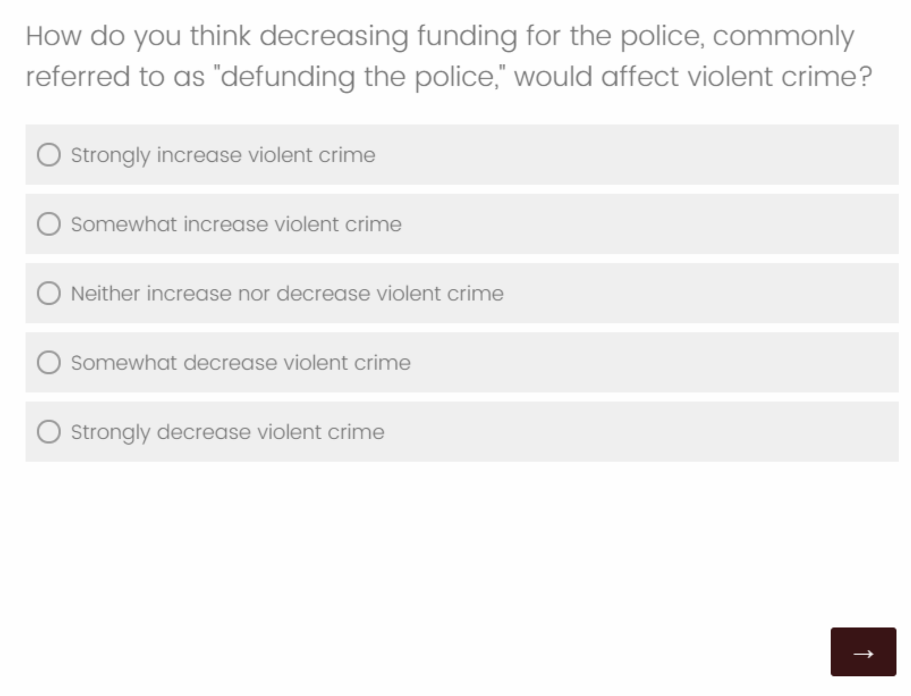
	Agree	Disagree
It is sometimes hard for me to go on with my work if I am not encouraged	○	○
I sometimes feel resentful when I don't get my way	○	○
On a few occasions, I have given up doing something because I thought too little of my ability	○	○
There have been times when I felt like rebelling against people in authority even though I knew they were right	○	○



No matter who I'm talking to, I'm always a good listener	○	○
There have been occasions when I took advantage of someone	○	○
I'm always willing to admit it when I make a mistake	○	○
I sometimes try to get even rather than forgive and forget	○	○
I am always courteous, even to people who are disagreeable	○	○
I have never been irked when people expressed ideas very different from my own	○	○
There have been times when I was quite jealous of the good fortune of others	○	○
I am sometimes irritated by people who ask favors of me	○	○
I have deliberately said something that hurt someone's feelings	○	○

E.6 Auxiliary Experiment 1: Persuasion experiment – Democrats

E.6.1 Pre-treatment beliefs



How do you think decreasing funding for the police, commonly referred to as "defunding the police," would affect violent crime?

○ Strongly increase violent crime

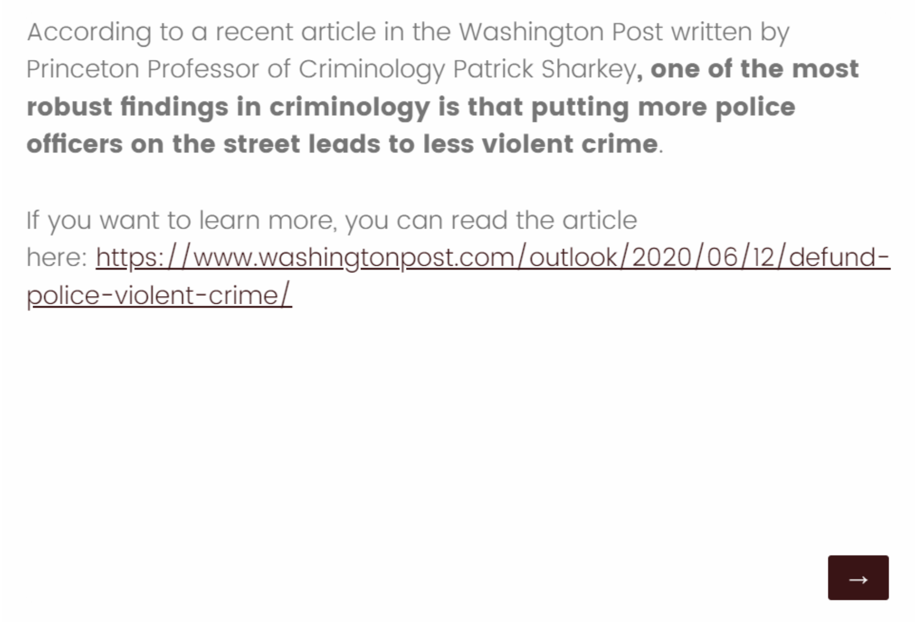
○ Somewhat increase violent crime

○ Neither increase nor decrease violent crime

○ Somewhat decrease violent crime

○ Strongly decrease violent crime

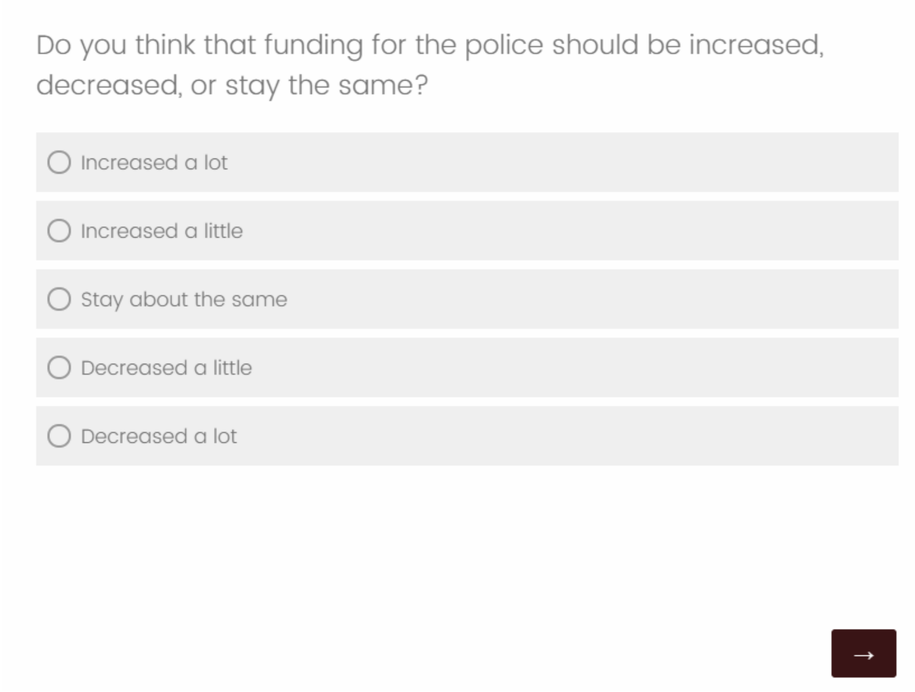
E.6.2 Information treatment (treatment group only)



According to a recent article in the Washington Post written by Princeton Professor of Criminology Patrick Sharkey, **one of the most robust findings in criminology is that putting more police officers on the street leads to less violent crime**.

If you want to learn more, you can read the article here: https://www.washingtonpost.com/outlook/2020/06/12/defund-police-violent-crime/

E.6.3 Post-treatment outcomes



Do you think that funding for the police should be increased, decreased, or stay the same?

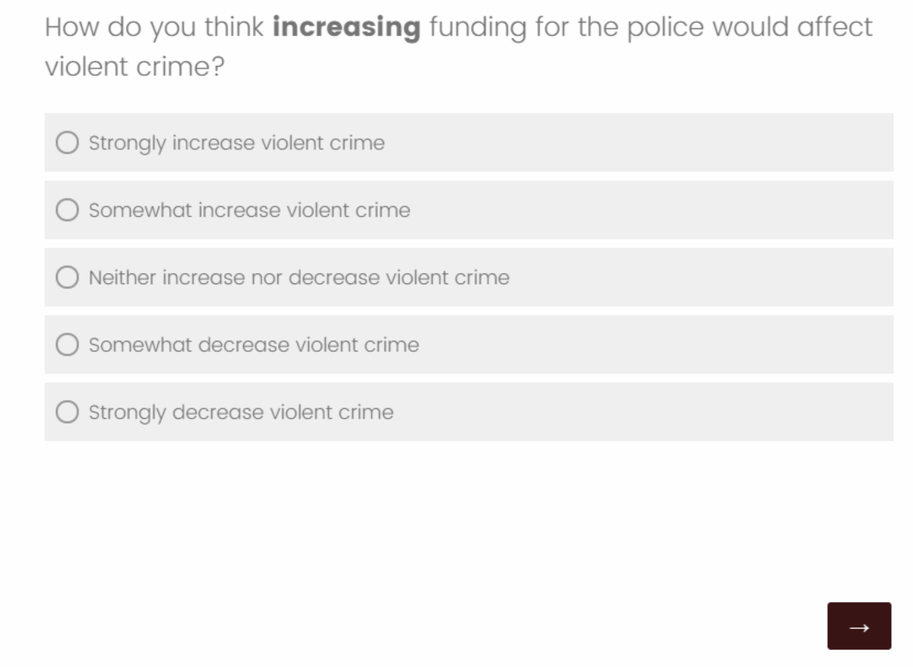
○ Increased a lot

○ Increased a little

○ Stay about the same

○ Decreased a little

○ Decreased a lot



How do you think **increasing** funding for the police would affect violent crime?

○ Strongly increase violent crime

○ Somewhat increase violent crime

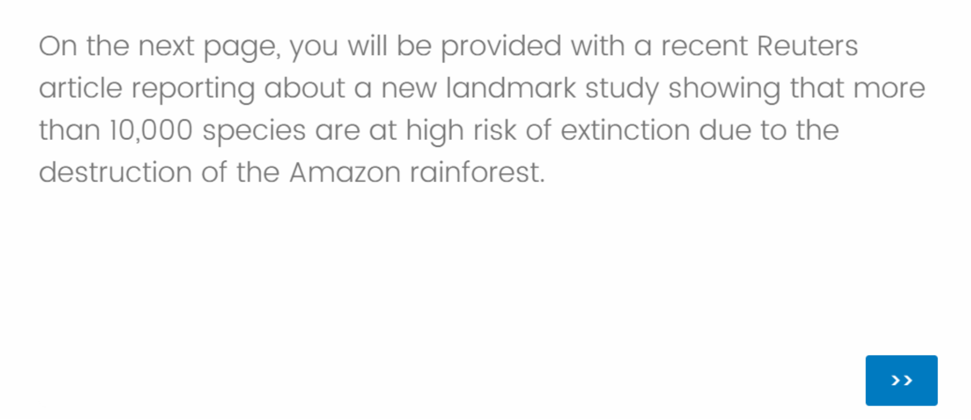
○ Neither increase nor decrease violent crime

○ Somewhat decrease violent crime

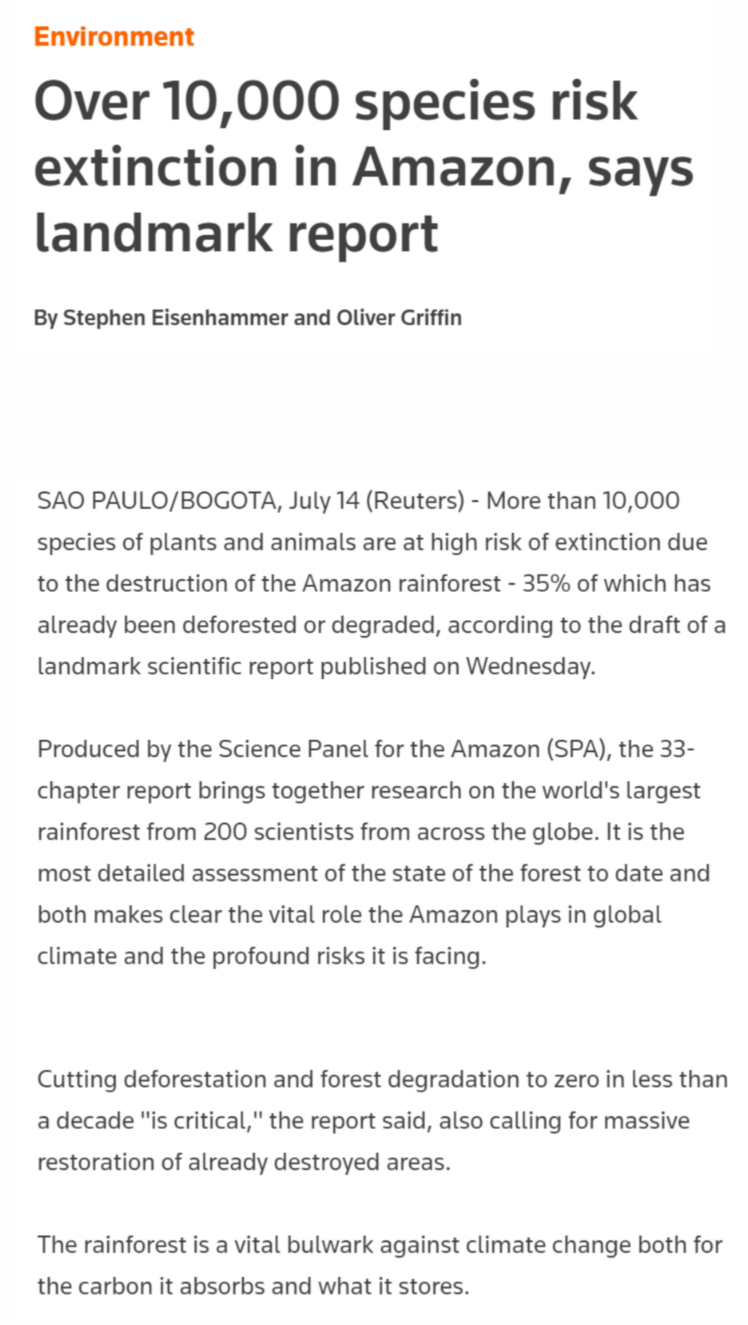
○ Strongly decrease violent crime

E.7 Auxiliary Experiment 2: Rainforest placebo

E.7.1 Pre-treatment questions



On the next page, you will be provided with a recent Reuters article reporting about a new landmark study showing that more than 10,000 species are at high risk of extinction due to the destruction of the Amazon rainforest.



Environment

Over 10,000 species risk extinction in Amazon, says landmark report

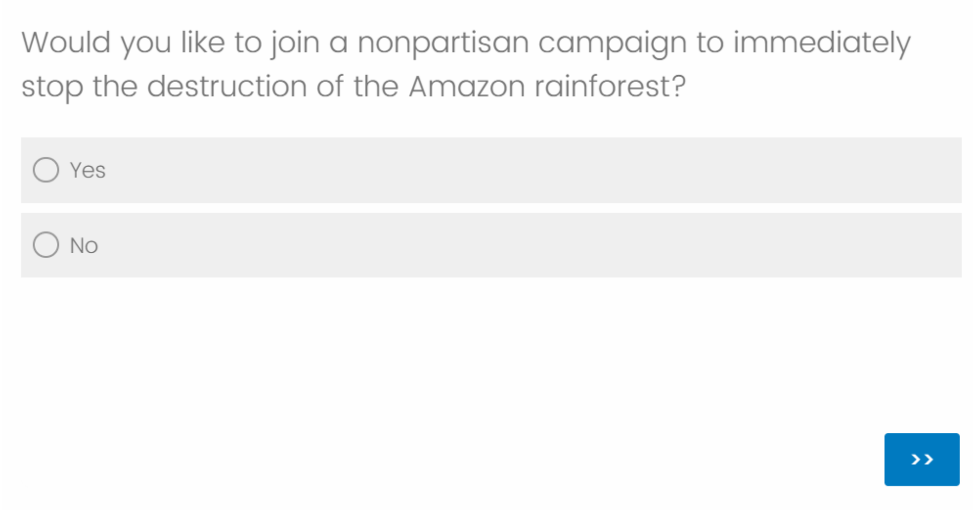
By Stephen Eisenhammer and Oliver Griffin

SAO PAULO/BOGOTA, July 14 (Reuters) - More than 10,000 species of plants and animals are at high risk of extinction due to the destruction of the Amazon rainforest - 35% of which has already been deforested or degraded, according to the draft of a landmark scientific report published on Wednesday.

Produced by the Science Panel for the Amazon (SPA), the 33-chapter report brings together research on the world's largest rainforest from 200 scientists from across the globe. It is the most detailed assessment of the state of the forest to date and both makes clear the vital role the Amazon plays in global climate and the profound risks it is facing.

Cutting deforestation and forest degradation to zero in less than a decade "is critical," the report said, also calling for massive restoration of already destroyed areas.

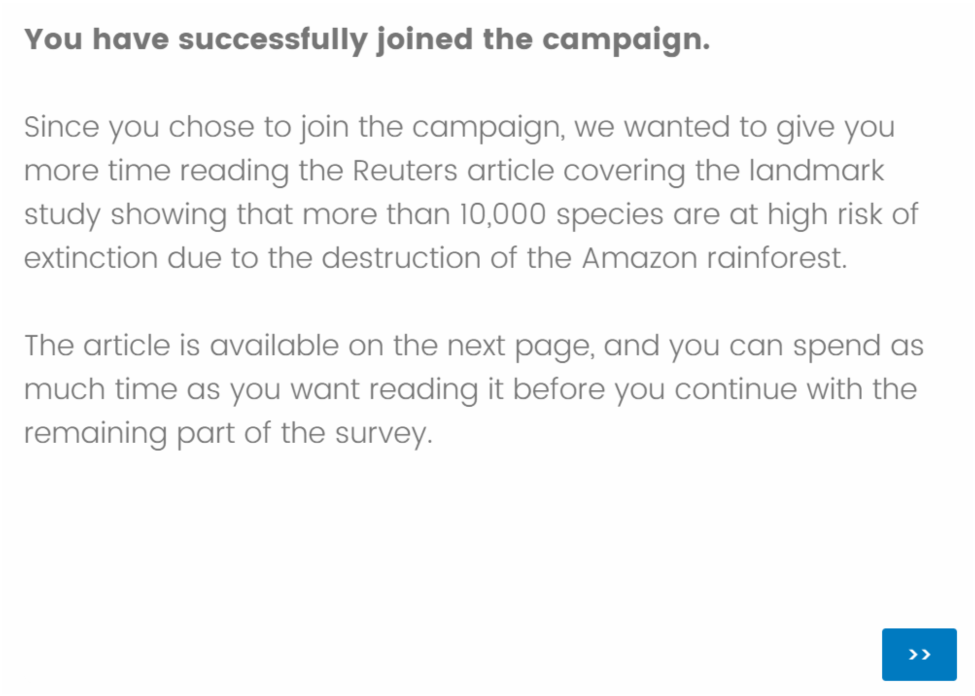
The rainforest is a vital bulwark against climate change both for the carbon it absorbs and what it stores.



Would you like to join a nonpartisan campaign to immediately stop the destruction of the Amazon rainforest?

○ Yes

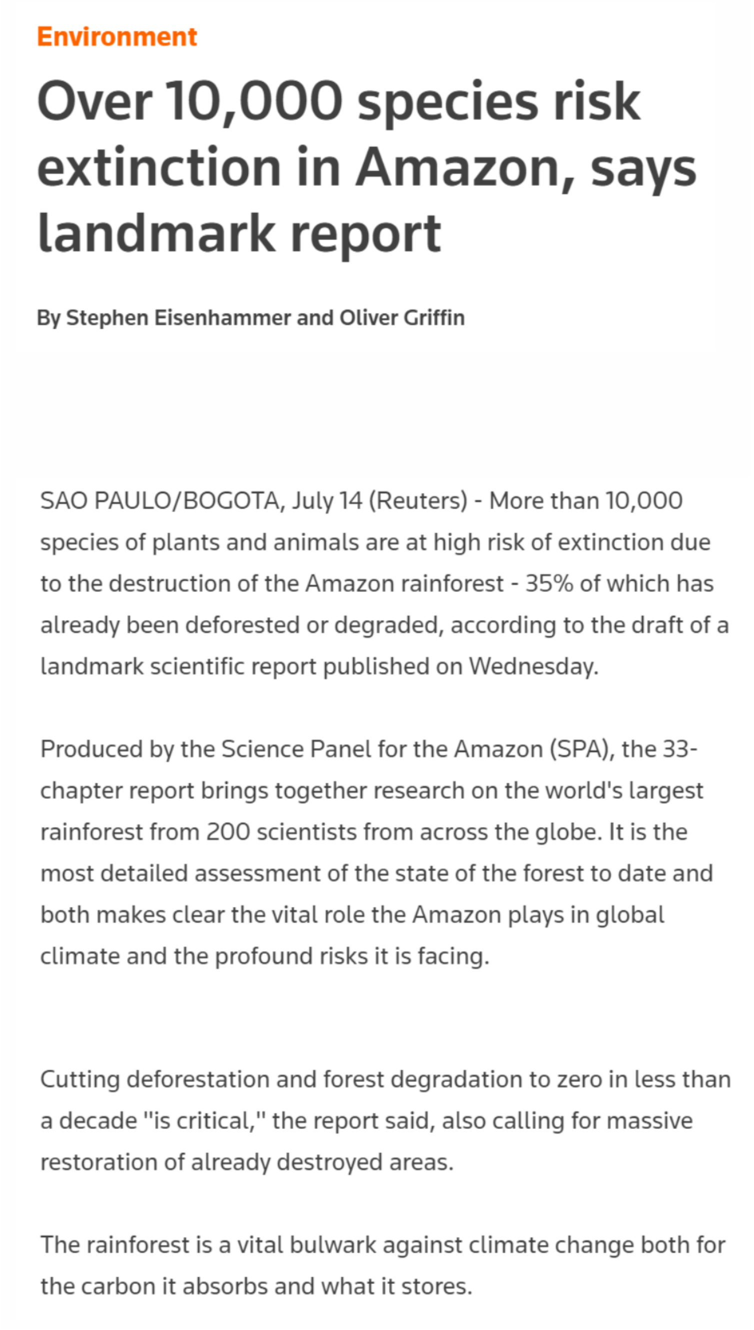
○ No



You have successfully joined the campaign.

Since you chose to join the campaign, we wanted to give you more time reading the Reuters article covering the landmark study showing that more than 10,000 species are at high risk of extinction due to the destruction of the Amazon rainforest.

The article is available on the next page, and you can spend as much time as you want reading it before you continue with the remaining part of the survey.



Environment

Over 10,000 species risk extinction in Amazon, says landmark report

By Stephen Eisenhammer and Oliver Griffin

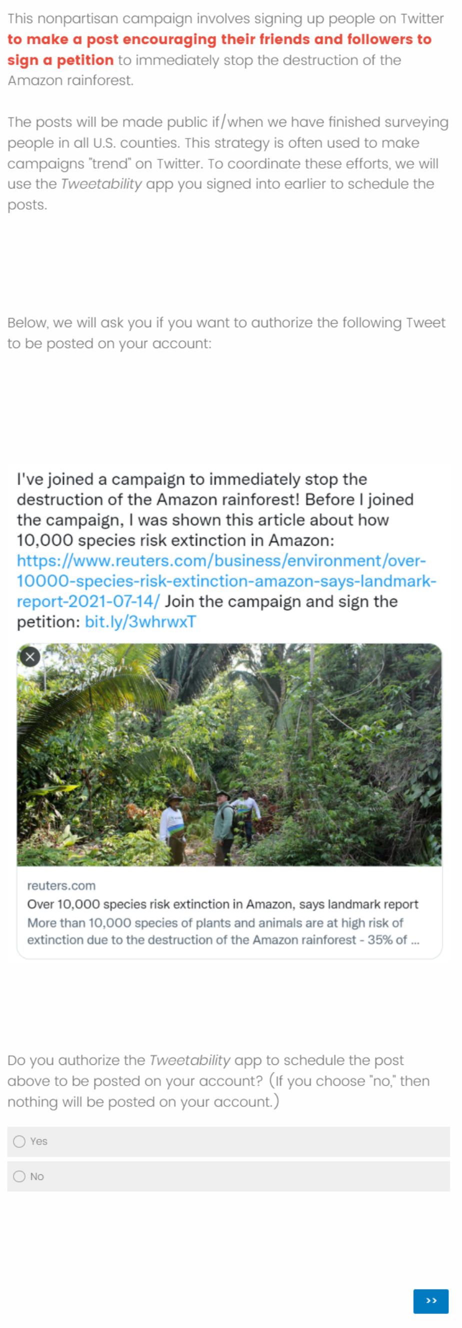
SAO PAULO/BOGOTA, July 14 (Reuters) - More than 10,000 species of plants and animals are at high risk of extinction due to the destruction of the Amazon rainforest - 35% of which has already been deforested or degraded, according to the draft of a landmark scientific report published on Wednesday.

Produced by the Science Panel for the Amazon (SPA), the 33-chapter report brings together research on the world's largest rainforest from 200 scientists from across the globe. It is the most detailed assessment of the state of the forest to date and both makes clear the vital role the Amazon plays in global climate and the profound risks it is facing.

Cutting deforestation and forest degradation to zero in less than a decade "is critical," the report said, also calling for massive restoration of already destroyed areas.

The rainforest is a vital bulwark against climate change both for the carbon it absorbs and what it stores.

E.7.2 Treatment: “Before” wording (rationale)



This nonpartisan campaign involves signing up people on Twitter **to make a post encouraging their friends and followers to sign a petition** to immediately stop the destruction of the Amazon rainforest.

The posts will be made public if/when we have finished surveying people in all U.S. counties. This strategy is often used to make campaigns "trend" on Twitter. To coordinate these efforts, we will use the *Tweetability* app you signed into earlier to schedule the posts.

Below, we will ask you if you want to authorize the following Tweet to be posted on your account:

I've joined a campaign to immediately stop the destruction of the Amazon rainforest! Before I joined the campaign, I was shown this article about how 10,000 species risk extinction in Amazon: https://www.reuters.com/business/environment/over-10000-species-risk-extinction-amazon-says-landmark-report-2021-07-14/ Join the campaign and sign the petition: bit.ly/3whrwxT

reuters.com

Over 10,000 species risk extinction in Amazon, says landmark report

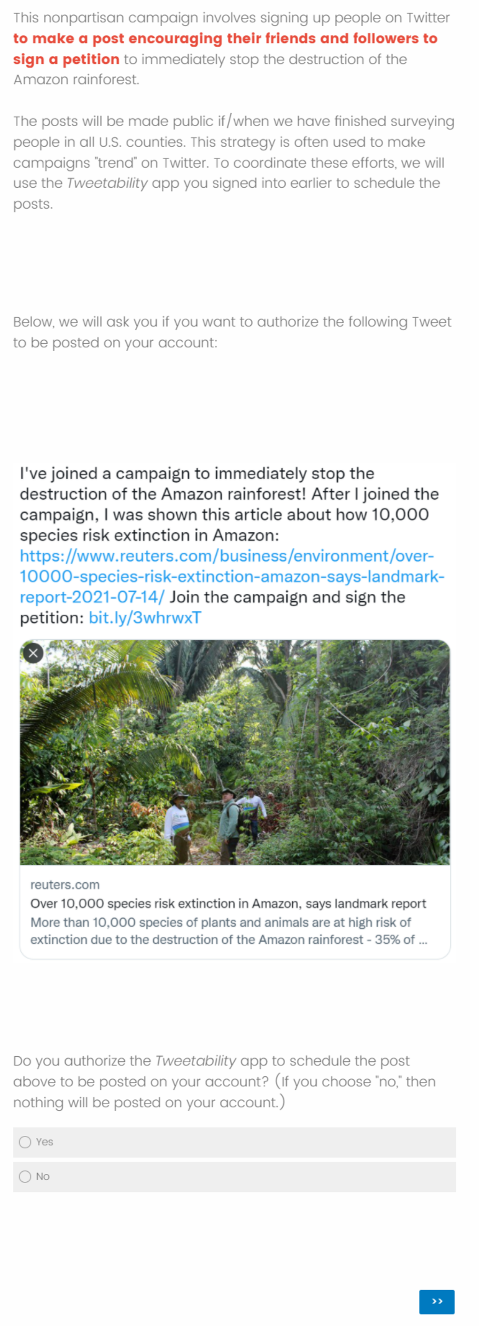
More than 10,000 species of plants and animals are at high risk of extinction due to the destruction of the Amazon rainforest - 35% of ...

Do you authorize the *Tweetability* app to schedule the post above to be posted on your account? (If you choose "no," then nothing will be posted on your account.)

○ Yes

○ No

E.7.3 Treatment: "After" wording (no rationale)



This nonpartisan campaign involves signing up people on Twitter **to make a post encouraging their friends and followers to sign a petition** to immediately stop the destruction of the Amazon rainforest.

The posts will be made public if/when we have finished surveying people in all U.S. counties. This strategy is often used to make campaigns "trend" on Twitter. To coordinate these efforts, we will use the *Tweetability* app you signed into earlier to schedule the posts.

Below, we will ask you if you want to authorize the following Tweet to be posted on your account:

I've joined a campaign to immediately stop the destruction of the Amazon rainforest! After I joined the campaign, I was shown this article about how 10,000 species risk extinction in Amazon: https://www.reuters.com/business/environment/over-10000-species-risk-extinction-amazon-says-landmark-report-2021-07-14/ Join the campaign and sign the petition: bit.ly/3whrwxT

reuters.com

Over 10,000 species risk extinction in Amazon, says landmark report

More than 10,000 species of plants and animals are at high risk of extinction due to the destruction of the Amazon rainforest - 35% of ...

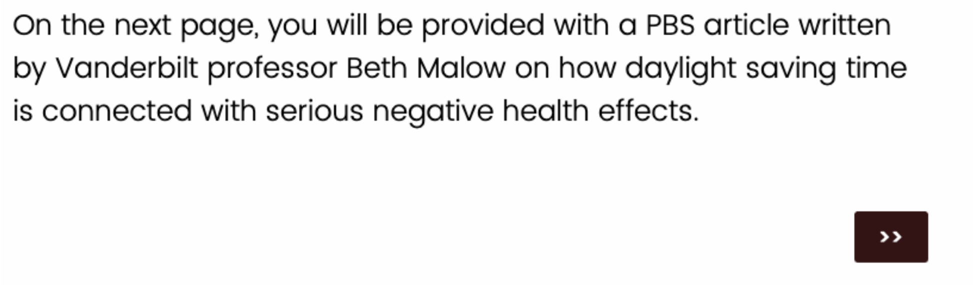
Do you authorize the *Tweetability* app to schedule the post above to be posted on your account? (If you choose "no," then nothing will be posted on your account.)

○ Yes

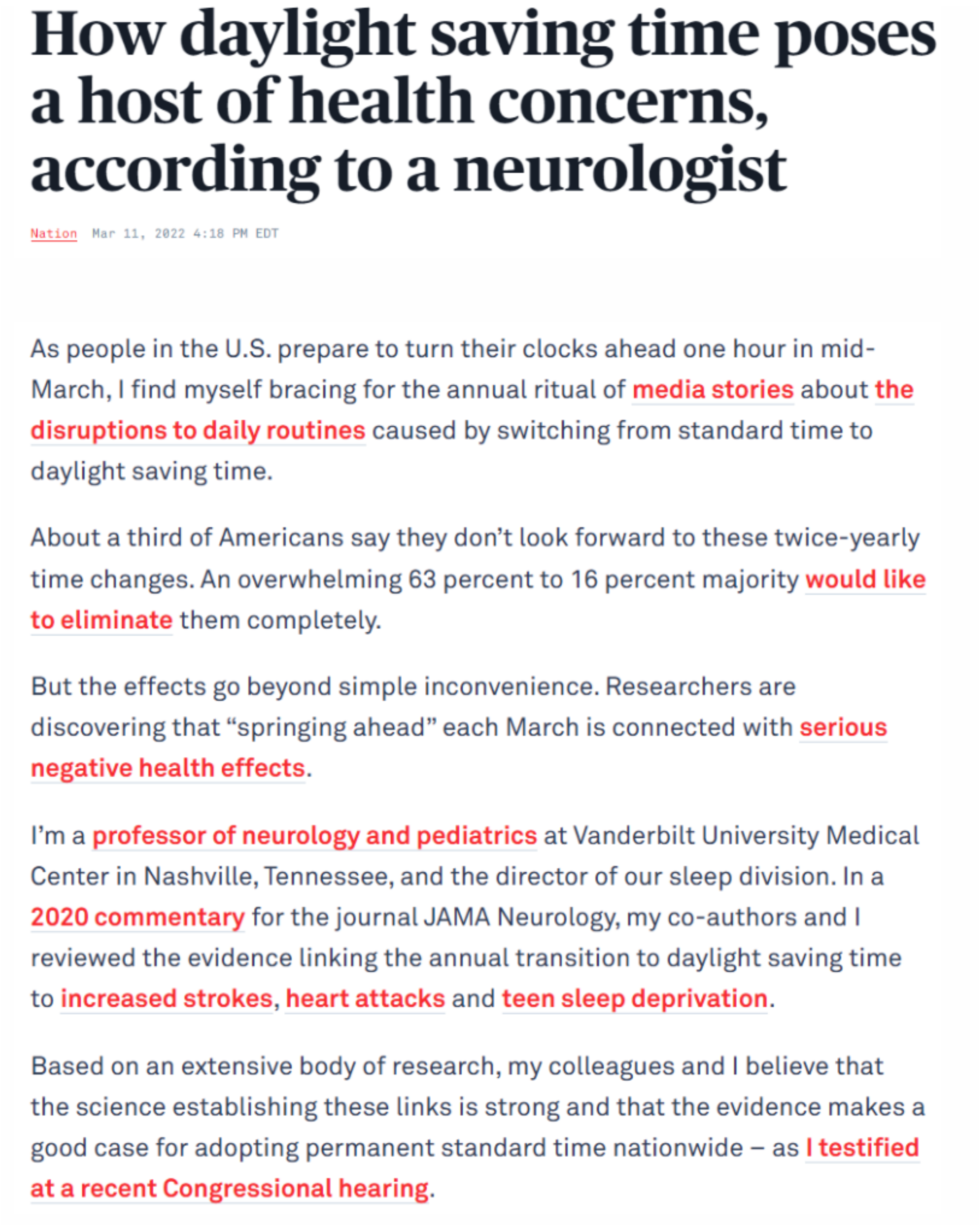
○ No

E.8 Auxiliary Experiment 3: Daylight Saving placebo

E.8.1 Pre-treatment questions



On the next page, you will be provided with a PBS article written by Vanderbilt professor Beth Malow on how daylight saving time is connected with serious negative health effects.



How daylight saving time poses a host of health concerns, according to a neurologist

Nation Mar 11, 2022 4:18 PM EDT

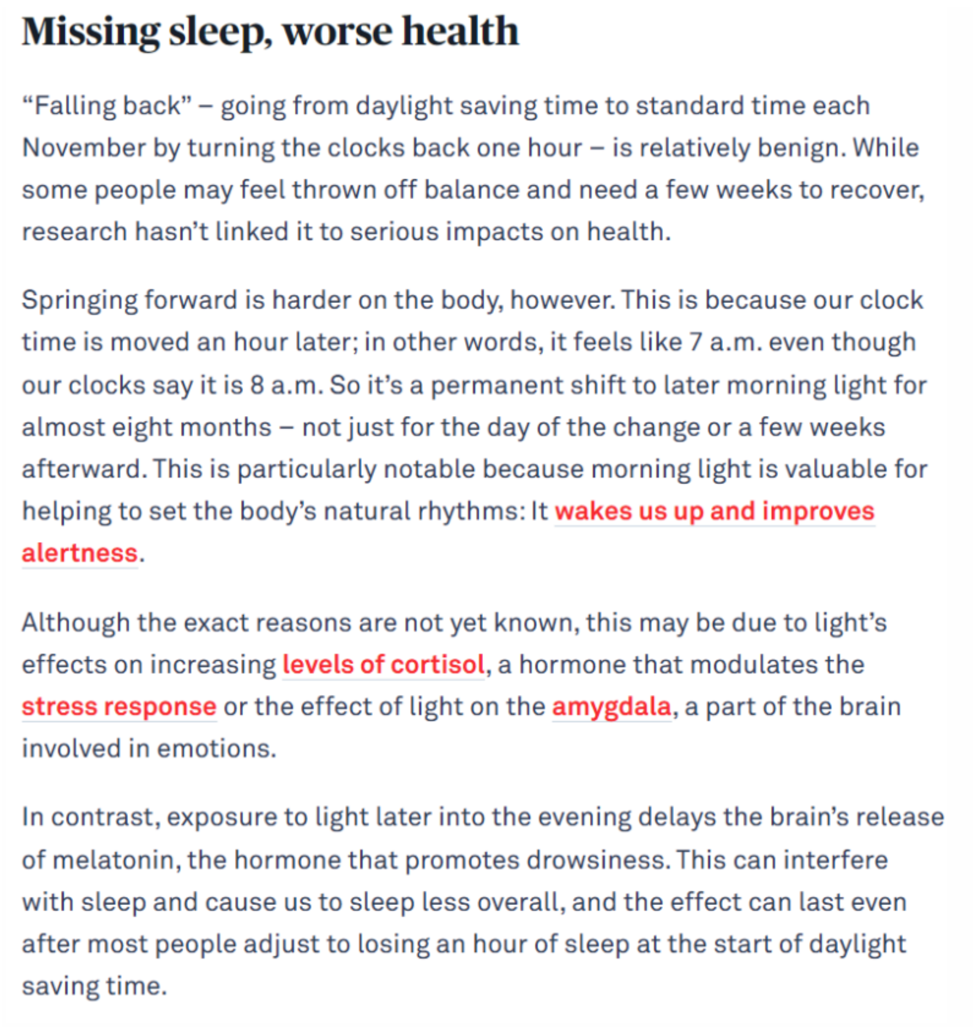
As people in the U.S. prepare to turn their clocks ahead one hour in mid-March, I find myself bracing for the annual ritual of **media stories** about **the disruptions to daily routines** caused by switching from standard time to daylight saving time.

About a third of Americans say they don't look forward to these twice-yearly time changes. An overwhelming 63 percent to 16 percent majority **would like to eliminate** them completely.

But the effects go beyond simple inconvenience. Researchers are discovering that "springing ahead" each March is connected with **serious negative health effects**.

I'm a **professor of neurology and pediatrics** at Vanderbilt University Medical Center in Nashville, Tennessee, and the director of our sleep division. In a **2020 commentary** for the journal JAMA Neurology, my co-authors and I reviewed the evidence linking the annual transition to daylight saving time to **increased strokes**, **heart attacks** and **teen sleep deprivation**.

Based on an extensive body of research, my colleagues and I believe that the science establishing these links is strong and that the evidence makes a good case for adopting permanent standard time nationwide – as **I testified at a recent Congressional hearing**.



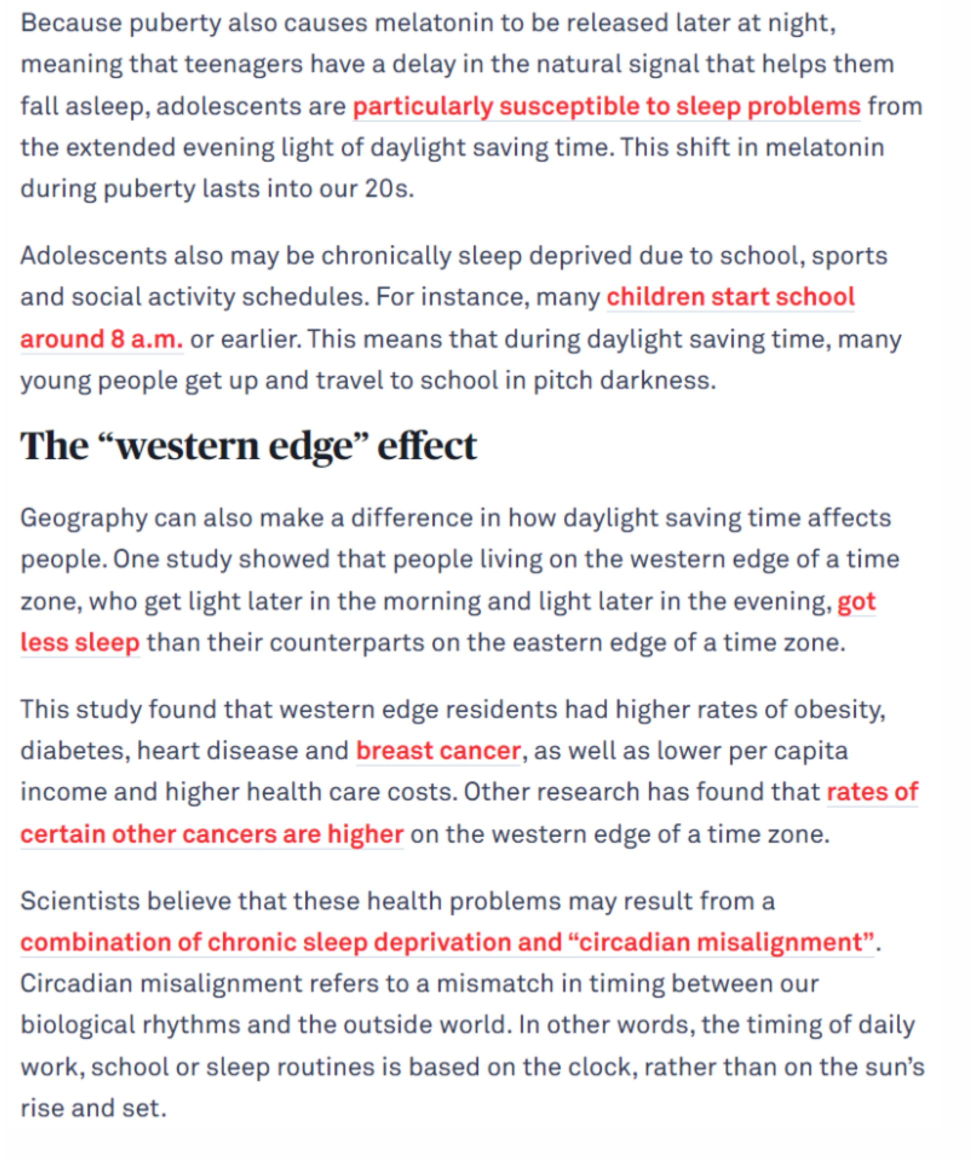
Missing sleep, worse health

“Falling back” – going from daylight saving time to standard time each November by turning the clocks back one hour – is relatively benign. While some people may feel thrown off balance and need a few weeks to recover, research hasn’t linked it to serious impacts on health.

Springing forward is harder on the body, however. This is because our clock time is moved an hour later; in other words, it feels like 7 a.m. even though our clocks say it is 8 a.m. So it’s a permanent shift to later morning light for almost eight months – not just for the day of the change or a few weeks afterward. This is particularly notable because morning light is valuable for helping to set the body’s natural rhythms: It **wakes us up and improves alertness**.

Although the exact reasons are not yet known, this may be due to light’s effects on increasing **levels of cortisol**, a hormone that modulates the **stress response** or the effect of light on the **amygdala**, a part of the brain involved in emotions.

In contrast, exposure to light later into the evening delays the brain’s release of melatonin, the hormone that promotes drowsiness. This can interfere with sleep and cause us to sleep less overall, and the effect can last even after most people adjust to losing an hour of sleep at the start of daylight saving time.



Because puberty also causes melatonin to be released later at night, meaning that teenagers have a delay in the natural signal that helps them fall asleep, adolescents are **particularly susceptible to sleep problems** from the extended evening light of daylight saving time. This shift in melatonin during puberty lasts into our 20s.

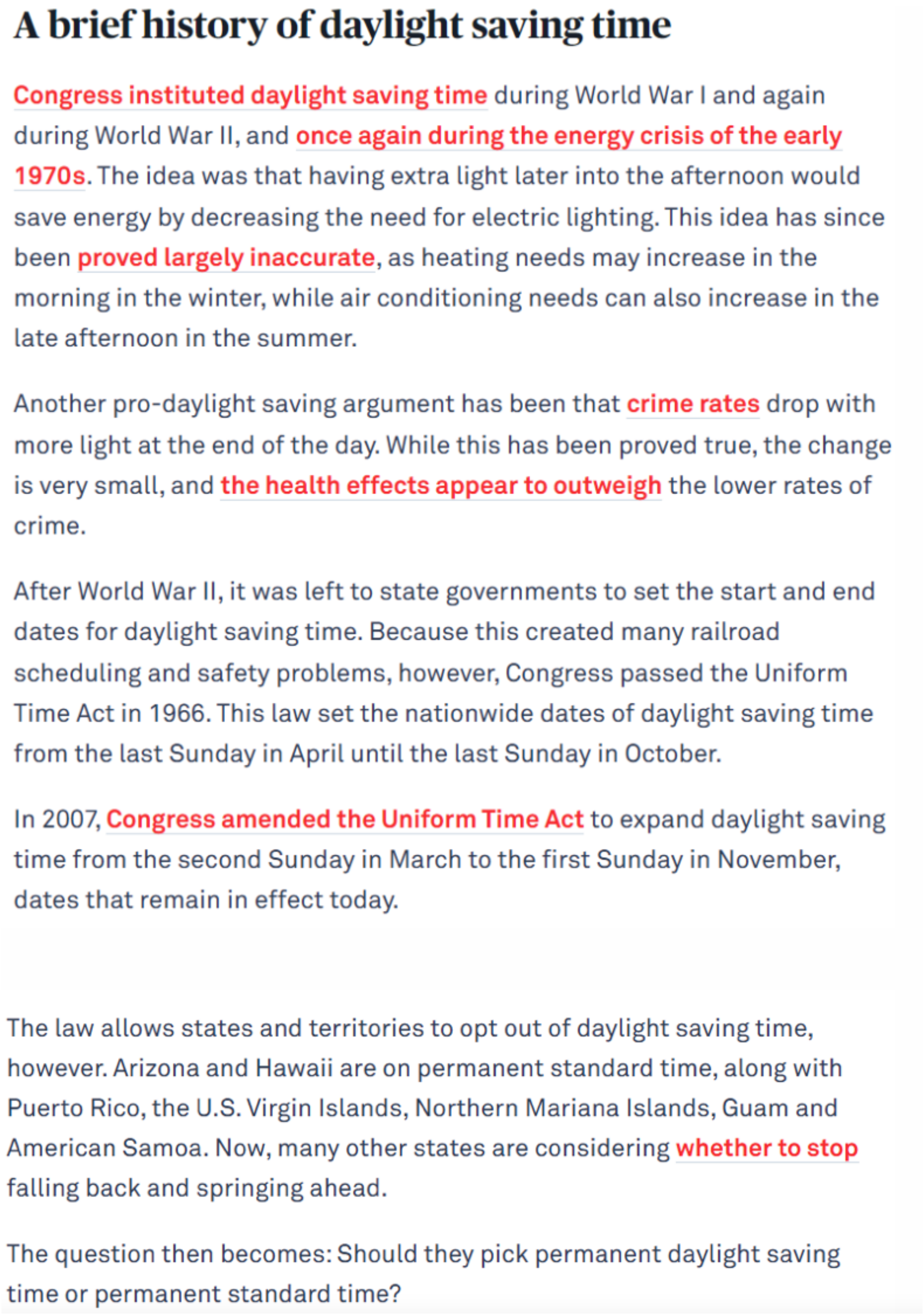
Adolescents also may be chronically sleep deprived due to school, sports and social activity schedules. For instance, many **children start school around 8 a.m.** or earlier. This means that during daylight saving time, many young people get up and travel to school in pitch darkness.

The “western edge” effect

Geography can also make a difference in how daylight saving time affects people. One study showed that people living on the western edge of a time zone, who get light later in the morning and light later in the evening, **got less sleep** than their counterparts on the eastern edge of a time zone.

This study found that western edge residents had higher rates of obesity, diabetes, heart disease and **breast cancer**, as well as lower per capita income and higher health care costs. Other research has found that **rates of certain other cancers are higher** on the western edge of a time zone.

Scientists believe that these health problems may result from a **combination of chronic sleep deprivation and “circadian misalignment”**. Circadian misalignment refers to a mismatch in timing between our biological rhythms and the outside world. In other words, the timing of daily work, school or sleep routines is based on the clock, rather than on the sun’s rise and set.



A brief history of daylight saving time

Congress instituted daylight saving time during World War I and again during World War II, and once again during the energy crisis of the early 1970s. The idea was that having extra light later into the afternoon would save energy by decreasing the need for electric lighting. This idea has since been proved largely inaccurate, as heating needs may increase in the morning in the winter, while air conditioning needs can also increase in the late afternoon in the summer.

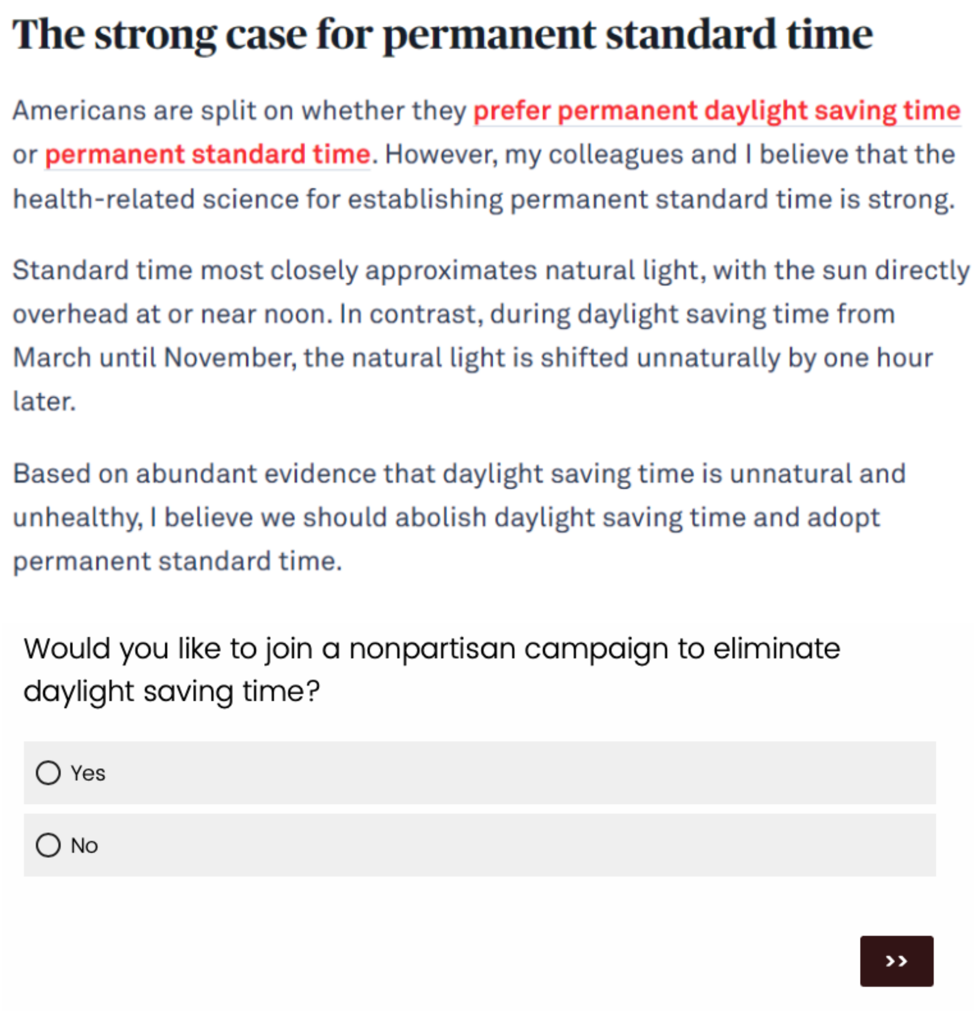
Another pro-daylight saving argument has been that crime rates drop with more light at the end of the day. While this has been proved true, the change is very small, and the health effects appear to outweigh the lower rates of crime.

After World War II, it was left to state governments to set the start and end dates for daylight saving time. Because this created many railroad scheduling and safety problems, however, Congress passed the Uniform Time Act in 1966. This law set the nationwide dates of daylight saving time from the last Sunday in April until the last Sunday in October.

In 2007, Congress amended the Uniform Time Act to expand daylight saving time from the second Sunday in March to the first Sunday in November, dates that remain in effect today.

The law allows states and territories to opt out of daylight saving time, however. Arizona and Hawaii are on permanent standard time, along with Puerto Rico, the U.S. Virgin Islands, Northern Mariana Islands, Guam and American Samoa. Now, many other states are considering whether to stop falling back and springing ahead.

The question then becomes: Should they pick permanent daylight saving time or permanent standard time?



The strong case for permanent standard time

Americans are split on whether they **prefer permanent daylight saving time** or **permanent standard time**. However, my colleagues and I believe that the health-related science for establishing permanent standard time is strong.

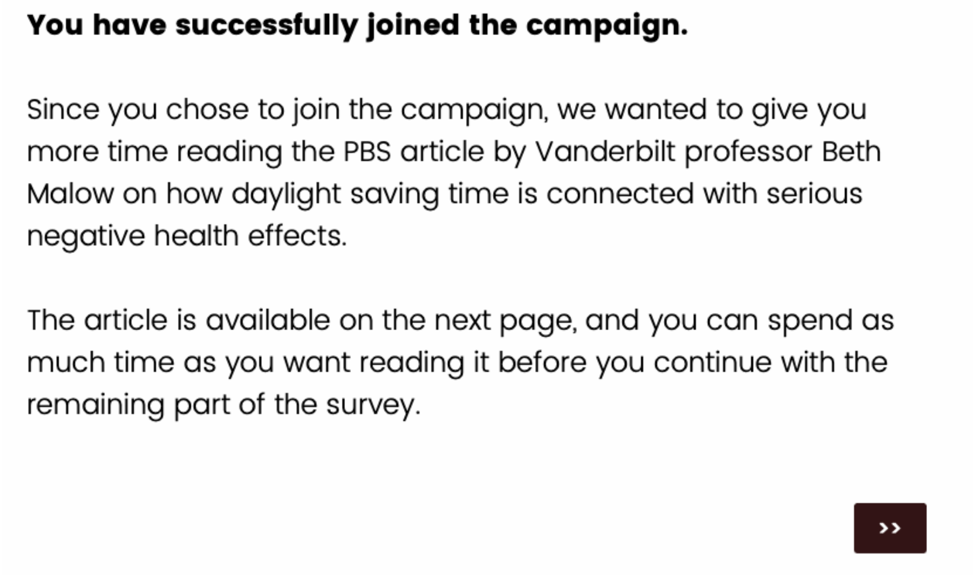
Standard time most closely approximates natural light, with the sun directly overhead at or near noon. In contrast, during daylight saving time from March until November, the natural light is shifted unnaturally by one hour later.

Based on abundant evidence that daylight saving time is unnatural and unhealthy, I believe we should abolish daylight saving time and adopt permanent standard time.

Would you like to join a nonpartisan campaign to eliminate daylight saving time?

○ Yes

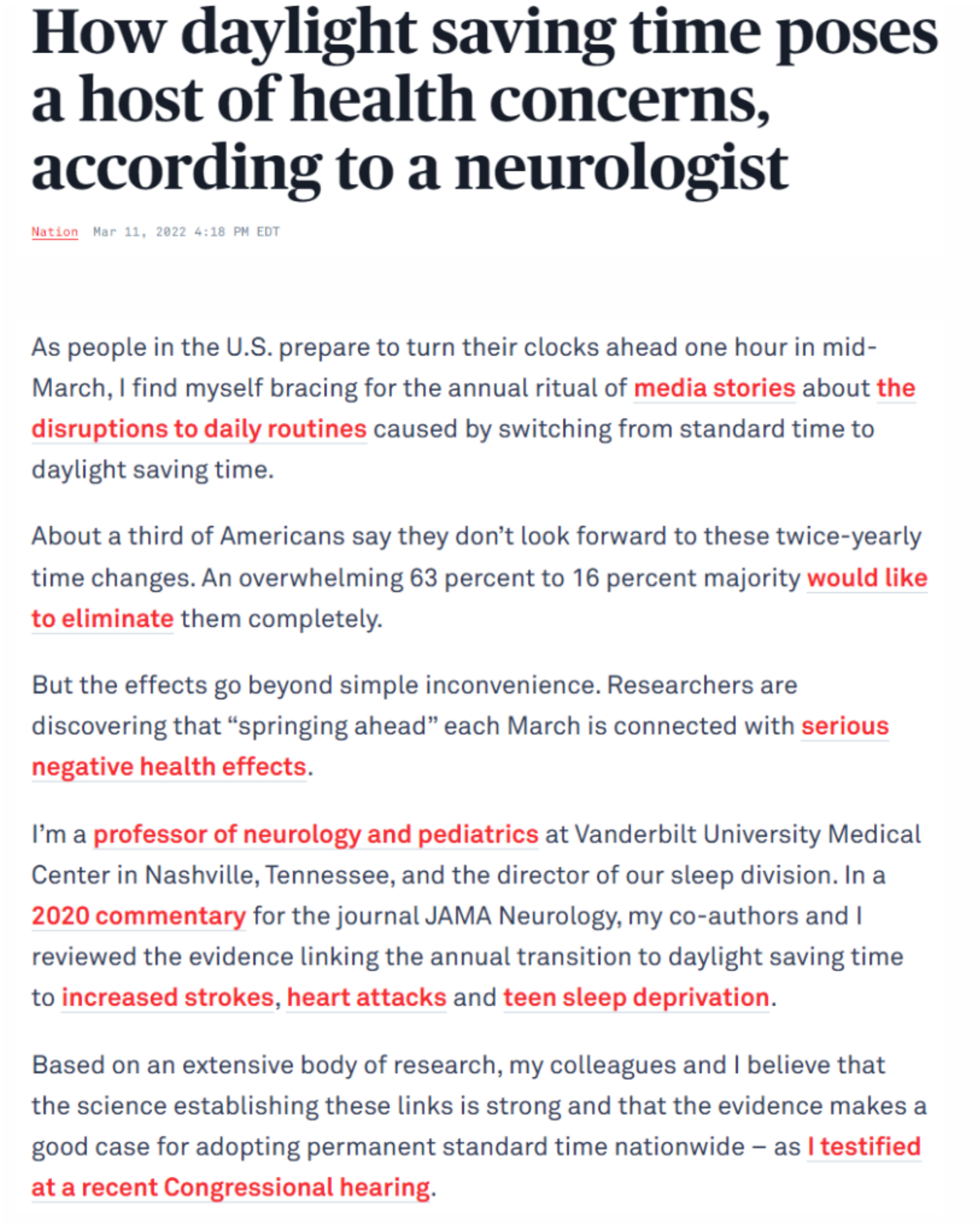
○ No



You have successfully joined the campaign.

Since you chose to join the campaign, we wanted to give you more time reading the PBS article by Vanderbilt professor Beth Malow on how daylight saving time is connected with serious negative health effects.

The article is available on the next page, and you can spend as much time as you want reading it before you continue with the remaining part of the survey.



How daylight saving time poses a host of health concerns, according to a neurologist

Nation Mar 11, 2022 4:18 PM EDT

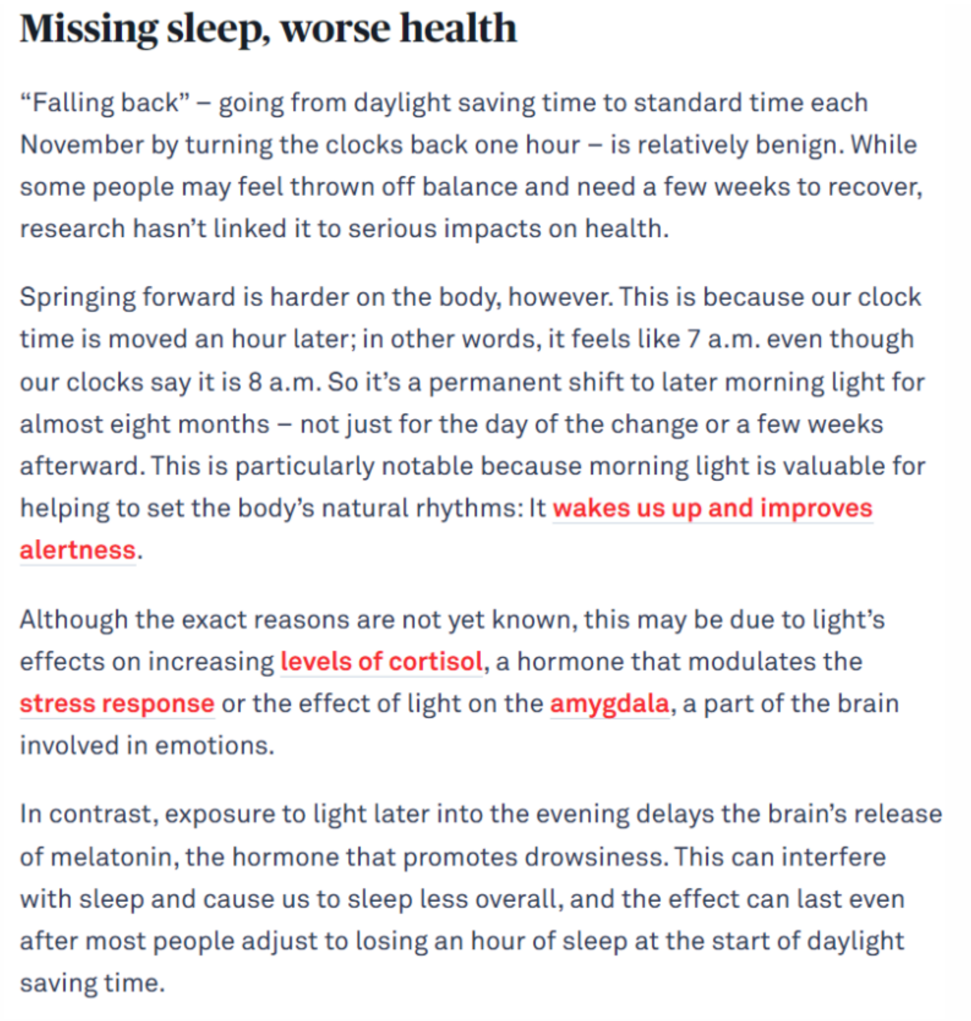
As people in the U.S. prepare to turn their clocks ahead one hour in mid-March, I find myself bracing for the annual ritual of media stories about the disruptions to daily routines caused by switching from standard time to daylight saving time.

About a third of Americans say they don't look forward to these twice-yearly time changes. An overwhelming 63 percent to 16 percent majority would like to eliminate them completely.

But the effects go beyond simple inconvenience. Researchers are discovering that "springing ahead" each March is connected with serious negative health effects.

I'm a professor of neurology and pediatrics at Vanderbilt University Medical Center in Nashville, Tennessee, and the director of our sleep division. In a 2020 commentary for the journal JAMA Neurology, my co-authors and I reviewed the evidence linking the annual transition to daylight saving time to increased strokes, heart attacks and teen sleep deprivation.

Based on an extensive body of research, my colleagues and I believe that the science establishing these links is strong and that the evidence makes a good case for adopting permanent standard time nationwide – as I testified at a recent Congressional hearing.



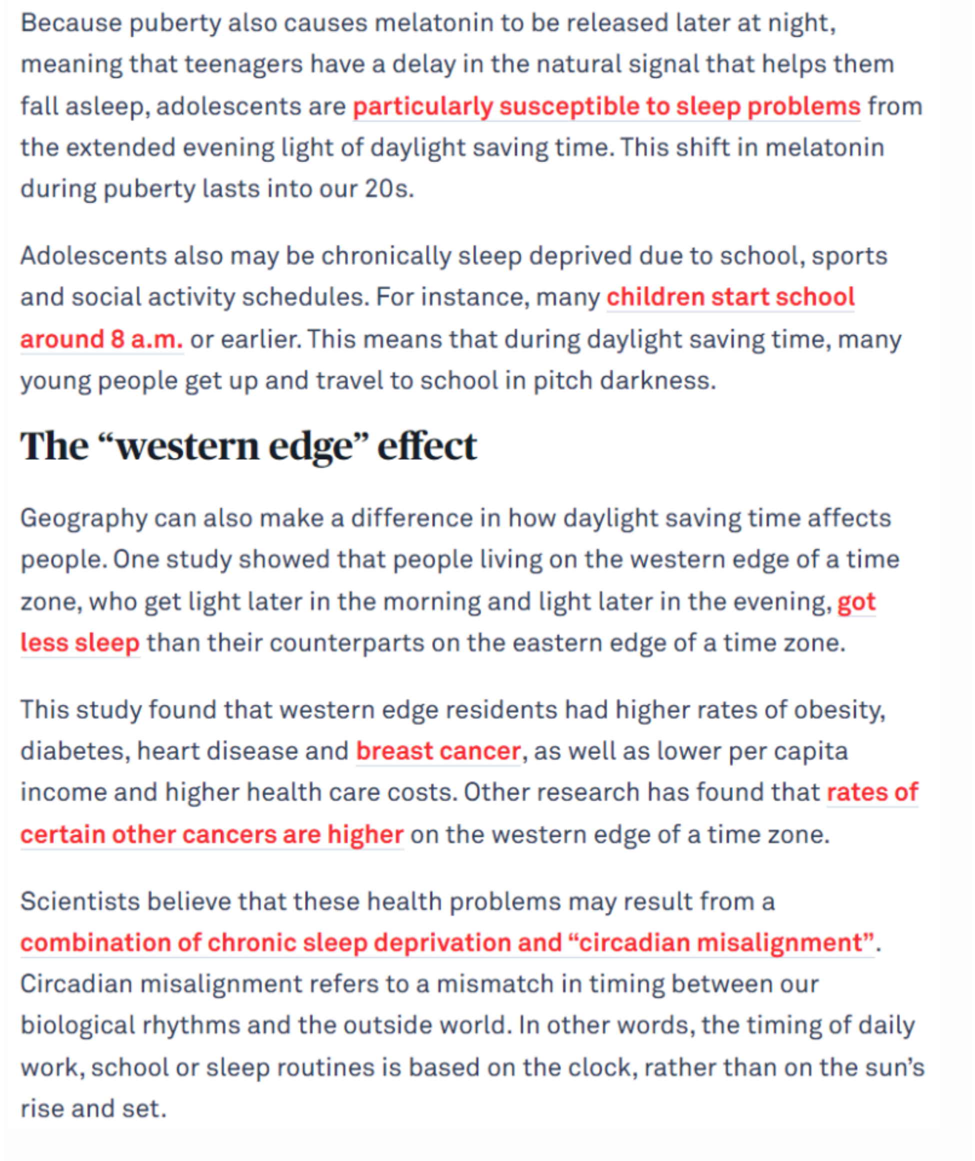
Missing sleep, worse health

“Falling back” – going from daylight saving time to standard time each November by turning the clocks back one hour – is relatively benign. While some people may feel thrown off balance and need a few weeks to recover, research hasn’t linked it to serious impacts on health.

Springing forward is harder on the body, however. This is because our clock time is moved an hour later; in other words, it feels like 7 a.m. even though our clocks say it is 8 a.m. So it’s a permanent shift to later morning light for almost eight months – not just for the day of the change or a few weeks afterward. This is particularly notable because morning light is valuable for helping to set the body’s natural rhythms: It **wakes us up and improves alertness**.

Although the exact reasons are not yet known, this may be due to light’s effects on increasing **levels of cortisol**, a hormone that modulates the **stress response** or the effect of light on the **amygdala**, a part of the brain involved in emotions.

In contrast, exposure to light later into the evening delays the brain’s release of melatonin, the hormone that promotes drowsiness. This can interfere with sleep and cause us to sleep less overall, and the effect can last even after most people adjust to losing an hour of sleep at the start of daylight saving time.



Because puberty also causes melatonin to be released later at night, meaning that teenagers have a delay in the natural signal that helps them fall asleep, adolescents are particularly susceptible to sleep problems from the extended evening light of daylight saving time. This shift in melatonin during puberty lasts into our 20s.

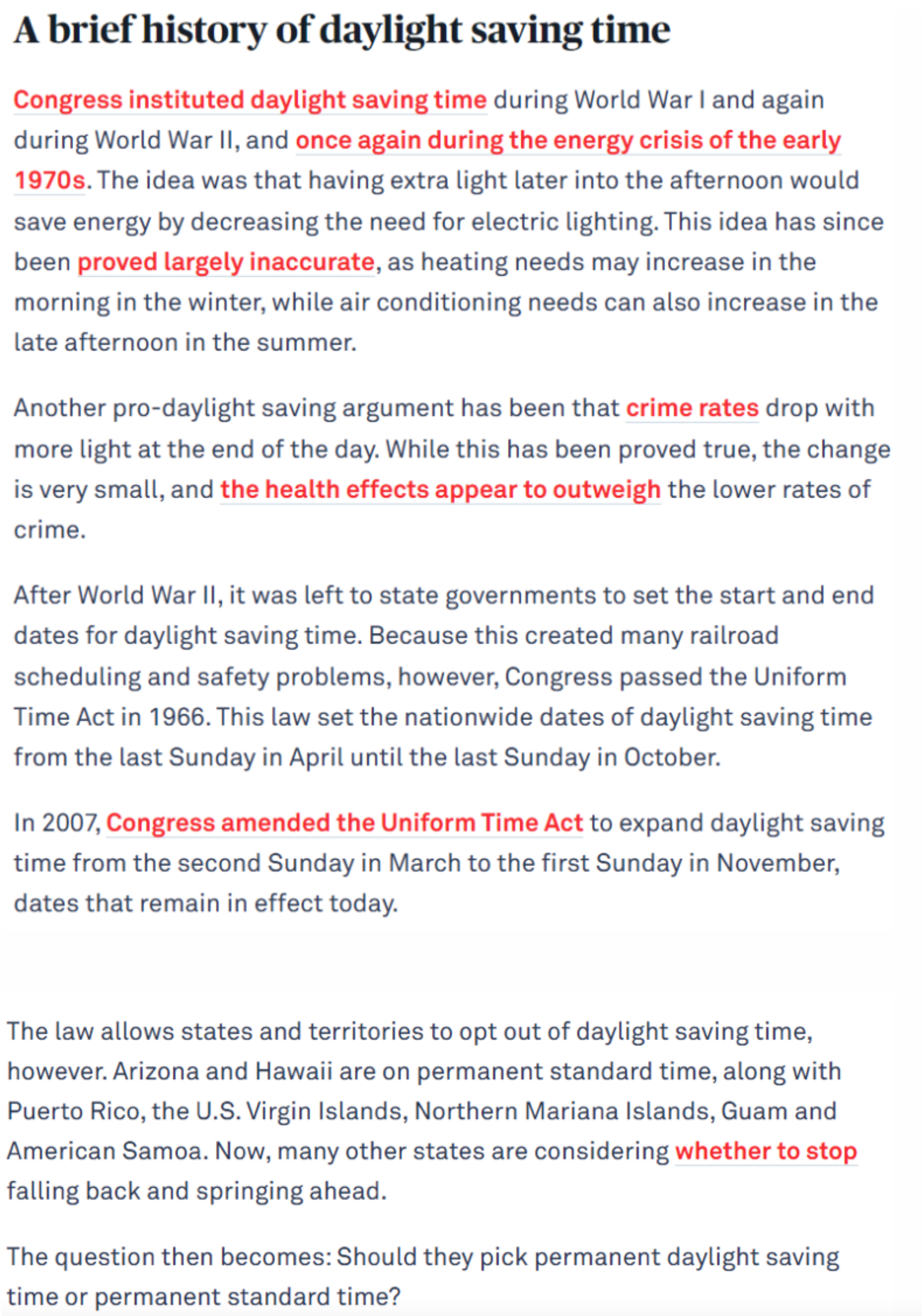
Adolescents also may be chronically sleep deprived due to school, sports and social activity schedules. For instance, many children start school around 8 a.m. or earlier. This means that during daylight saving time, many young people get up and travel to school in pitch darkness.

The “western edge” effect

Geography can also make a difference in how daylight saving time affects people. One study showed that people living on the western edge of a time zone, who get light later in the morning and light later in the evening, got less sleep than their counterparts on the eastern edge of a time zone.

This study found that western edge residents had higher rates of obesity, diabetes, heart disease and breast cancer, as well as lower per capita income and higher health care costs. Other research has found that rates of certain other cancers are higher on the western edge of a time zone.

Scientists believe that these health problems may result from a combination of chronic sleep deprivation and “circadian misalignment”. Circadian misalignment refers to a mismatch in timing between our biological rhythms and the outside world. In other words, the timing of daily work, school or sleep routines is based on the clock, rather than on the sun’s rise and set.



A brief history of daylight saving time

Congress instituted daylight saving time during World War I and again during World War II, and once again during the energy crisis of the early 1970s. The idea was that having extra light later into the afternoon would save energy by decreasing the need for electric lighting. This idea has since been proved largely inaccurate, as heating needs may increase in the morning in the winter, while air conditioning needs can also increase in the late afternoon in the summer.

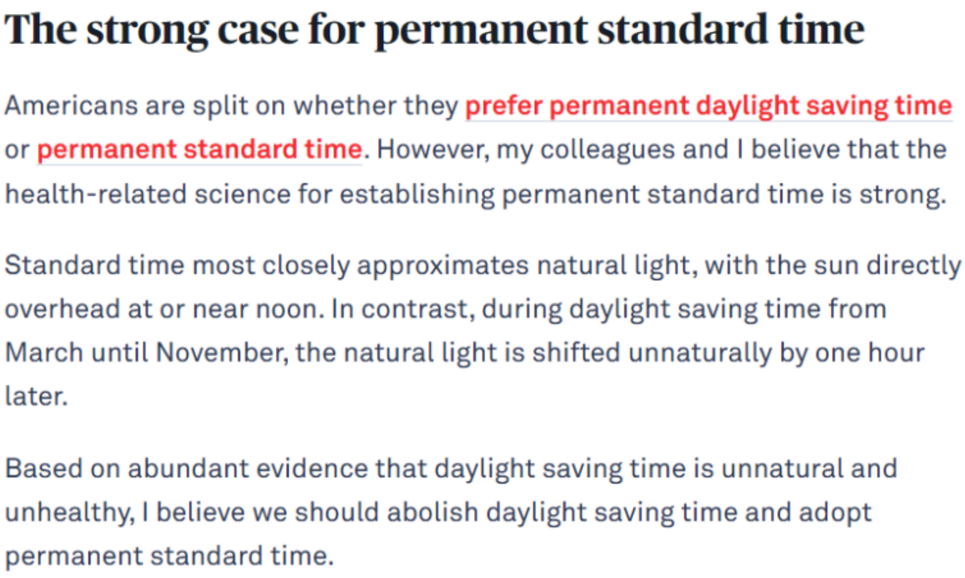
Another pro-daylight saving argument has been that crime rates drop with more light at the end of the day. While this has been proved true, the change is very small, and the health effects appear to outweigh the lower rates of crime.

After World War II, it was left to state governments to set the start and end dates for daylight saving time. Because this created many railroad scheduling and safety problems, however, Congress passed the Uniform Time Act in 1966. This law set the nationwide dates of daylight saving time from the last Sunday in April until the last Sunday in October.

In 2007, Congress amended the Uniform Time Act to expand daylight saving time from the second Sunday in March to the first Sunday in November, dates that remain in effect today.

The law allows states and territories to opt out of daylight saving time, however. Arizona and Hawaii are on permanent standard time, along with Puerto Rico, the U.S. Virgin Islands, Northern Mariana Islands, Guam and American Samoa. Now, many other states are considering whether to stop falling back and springing ahead.

The question then becomes: Should they pick permanent daylight saving time or permanent standard time?

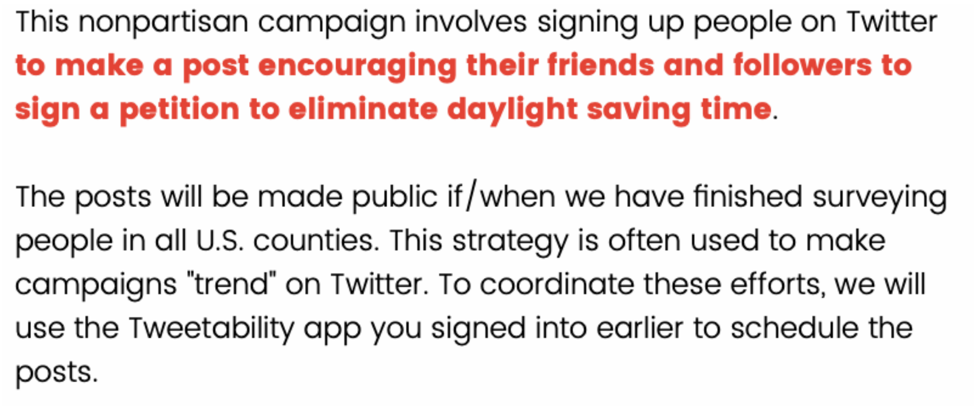


The strong case for permanent standard time

Americans are split on whether they **prefer permanent daylight saving time** or **permanent standard time**. However, my colleagues and I believe that the health-related science for establishing permanent standard time is strong.

Standard time most closely approximates natural light, with the sun directly overhead at or near noon. In contrast, during daylight saving time from March until November, the natural light is shifted unnaturally by one hour later.

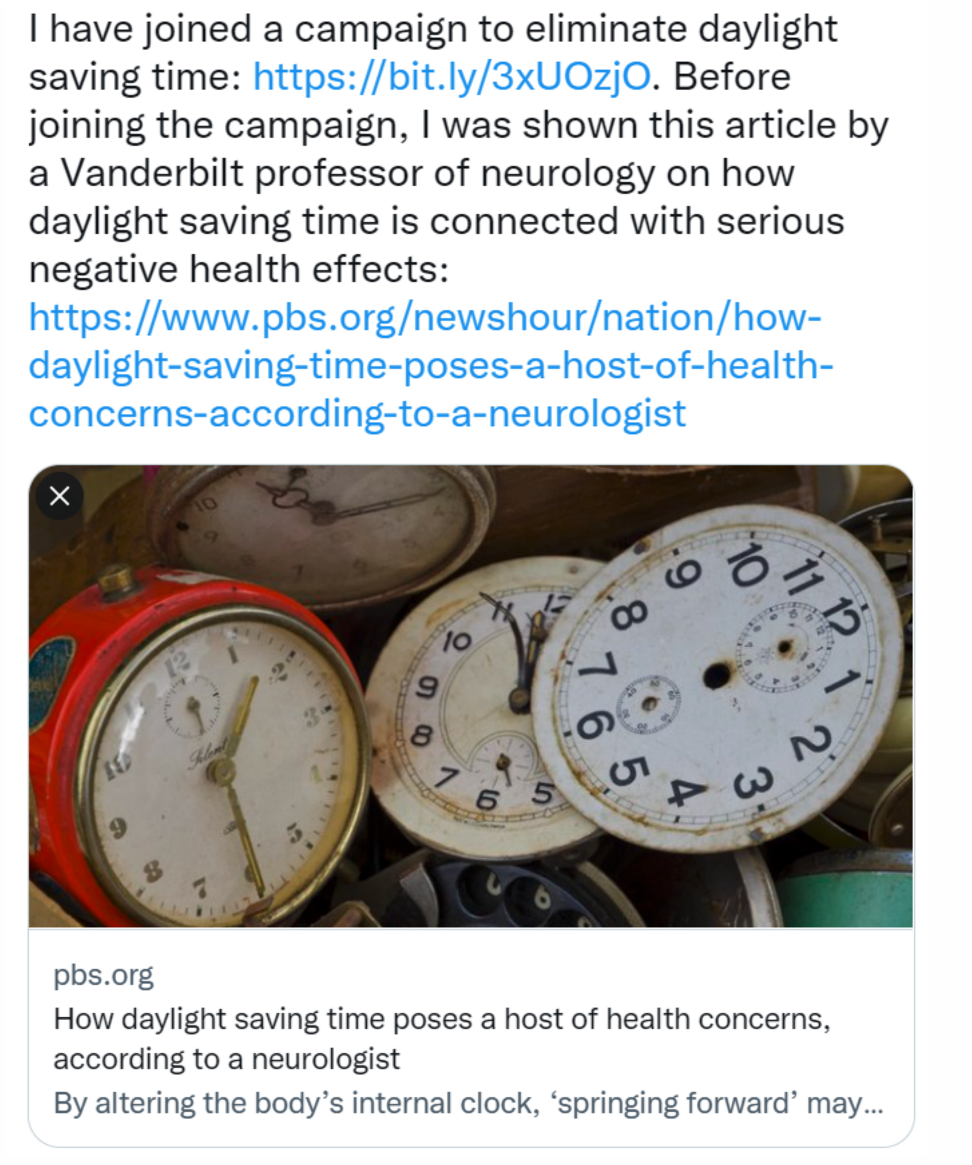
Based on abundant evidence that daylight saving time is unnatural and unhealthy, I believe we should abolish daylight saving time and adopt permanent standard time.



This nonpartisan campaign involves signing up people on Twitter **to make a post encouraging their friends and followers to sign a petition to eliminate daylight saving time**.

The posts will be made public if/when we have finished surveying people in all U.S. counties. This strategy is often used to make campaigns "trend" on Twitter. To coordinate these efforts, we will use the Tweetability app you signed into earlier to schedule the posts.

E.8.2 Treatment: “Before” wording (rationale)



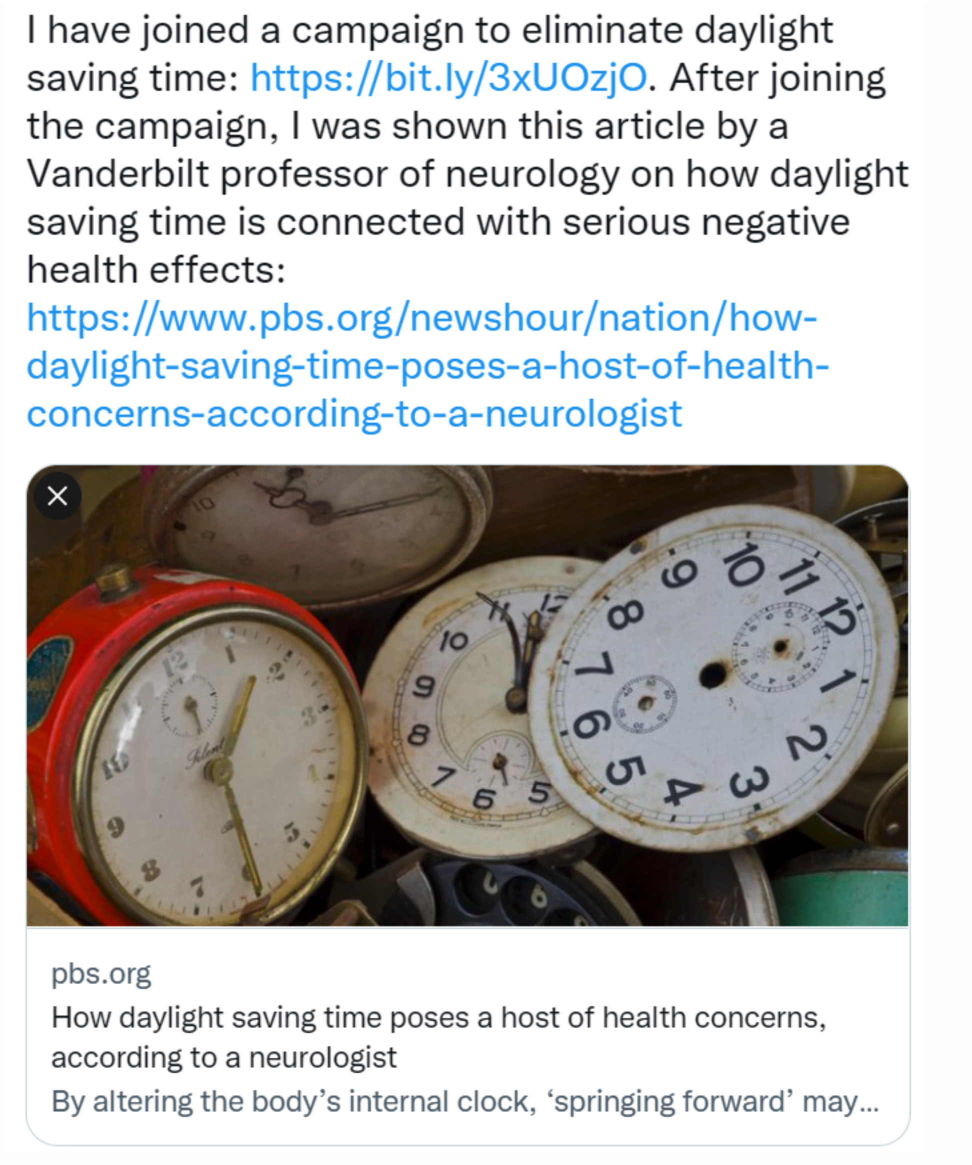
I have joined a campaign to eliminate daylight saving time: https://bit.ly/3xUOzjO. Before joining the campaign, I was shown this article by a Vanderbilt professor of neurology on how daylight saving time is connected with serious negative health effects: https://www.pbs.org/newshour/nation/how-daylight-saving-time-poses-a-host-of-health-concerns-according-to-a-neurologist

pbs.org

How daylight saving time poses a host of health concerns, according to a neurologist

By altering the body’s internal clock, ‘springing forward’ may...

E.8.3 Treatment: "After" wording (no rationale)



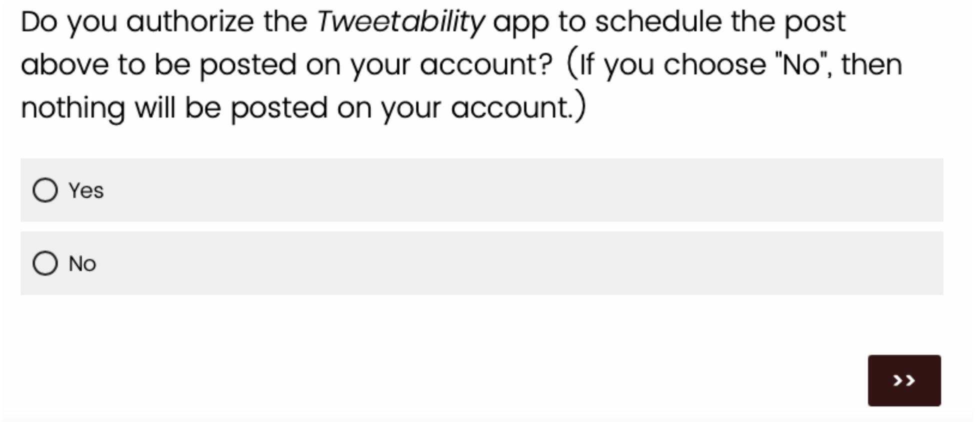
I have joined a campaign to eliminate daylight saving time: https://bit.ly/3xUOzjO. After joining the campaign, I was shown this article by a Vanderbilt professor of neurology on how daylight saving time is connected with serious negative health effects: https://www.pbs.org/newshour/nation/how-daylight-saving-time-poses-a-host-of-health-concerns-according-to-a-neurologist

pbs.org

How daylight saving time poses a host of health concerns, according to a neurologist

By altering the body's internal clock, 'springing forward' may...

E.9 Outcome

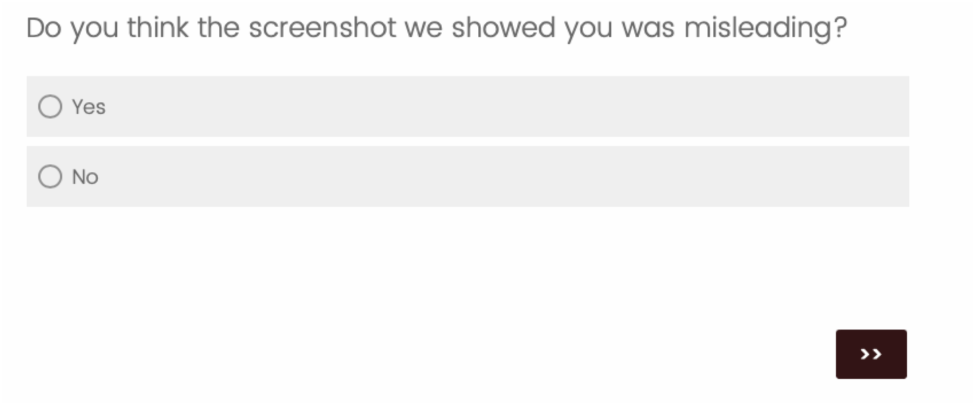


Do you authorize the *Tweetability* app to schedule the post above to be posted on your account? (If you choose "No", then nothing will be posted on your account.)

○ Yes

○ No

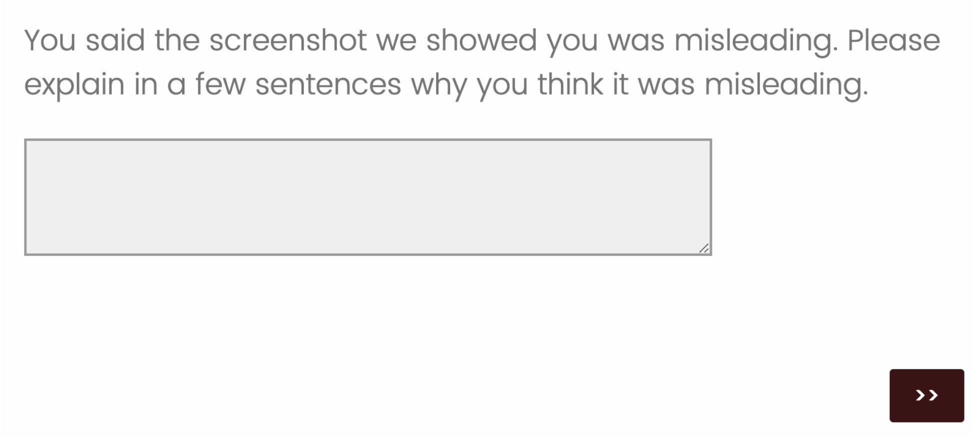
E.10 Misleading



Do you think the screenshot we showed you was misleading?

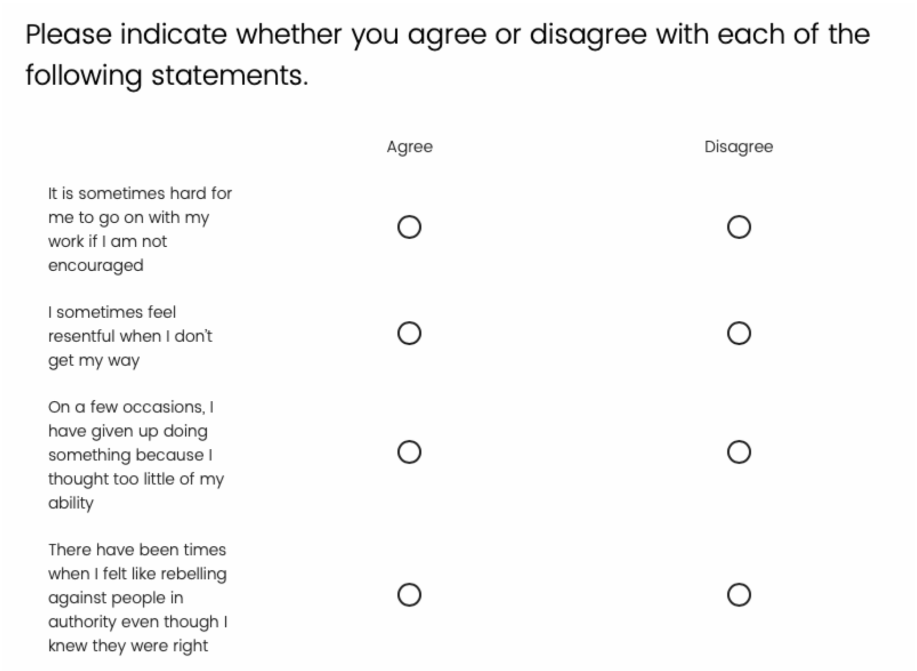
○ Yes

○ No



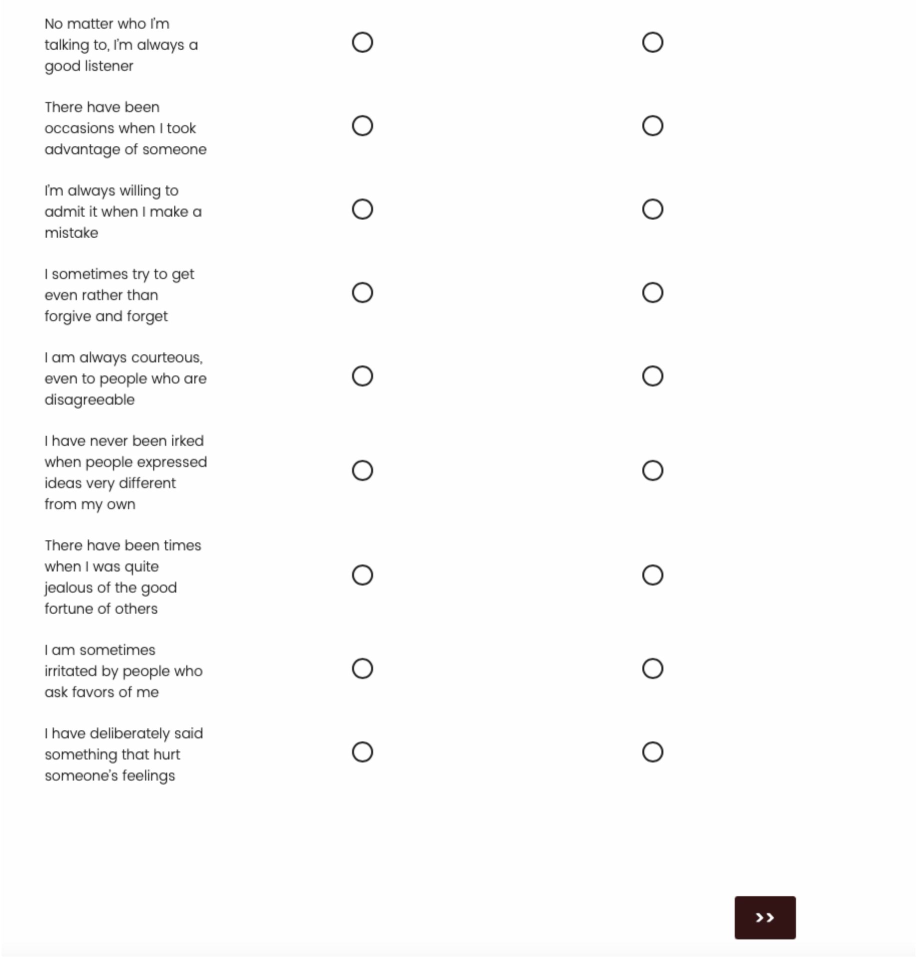
You said the screenshot we showed you was misleading. Please explain in a few sentences why you think it was misleading.

E.11 Marlowe-Crowne scale



Please indicate whether you agree or disagree with each of the following statements.

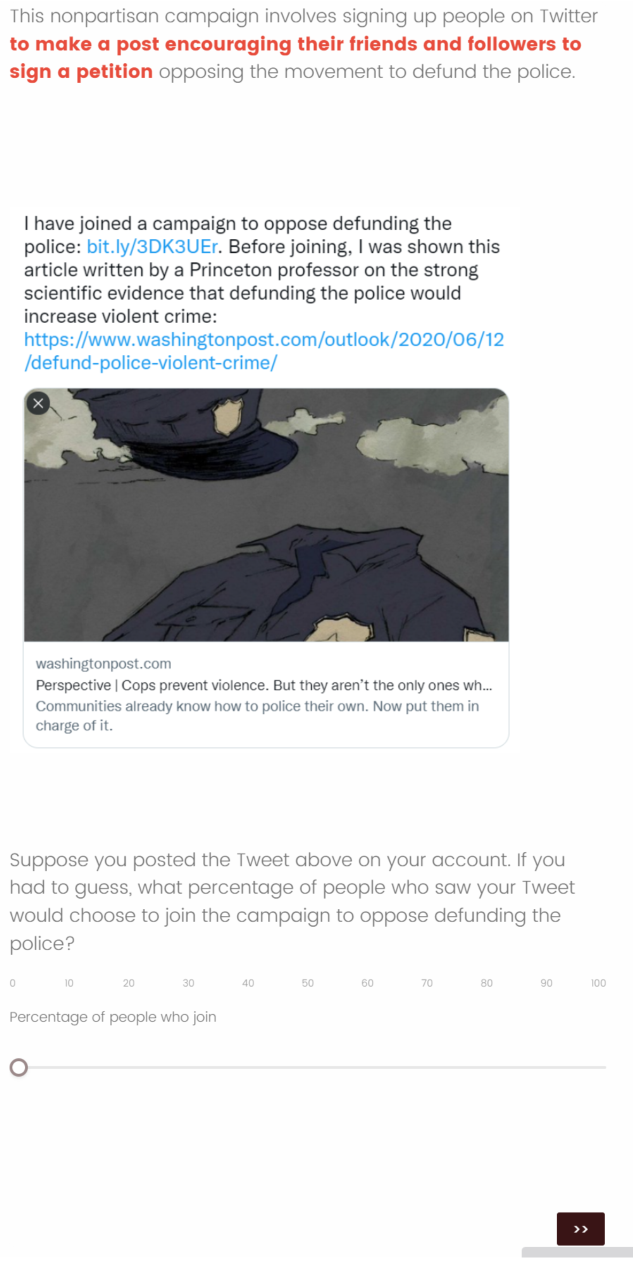
	Agree	Disagree
It is sometimes hard for me to go on with my work if I am not encouraged	○	○
I sometimes feel resentful when I don't get my way	○	○
On a few occasions, I have given up doing something because I thought too little of my ability	○	○
There have been times when I felt like rebelling against people in authority even though I knew they were right	○	○



No matter who I'm talking to, I'm always a good listener	○	○
There have been occasions when I took advantage of someone	○	○
I'm always willing to admit it when I make a mistake	○	○
I sometimes try to get even rather than forgive and forget	○	○
I am always courteous, even to people who are disagreeable	○	○
I have never been irked when people expressed ideas very different from my own	○	○
There have been times when I was quite jealous of the good fortune of others	○	○
I am sometimes irritated by people who ask favors of me	○	○
I have deliberately said something that hurt someone's feelings	○	○

E.12 Auxiliary Experiment 4: Anticipated persuasion – Democrats

E.12.1 Treatment: “Before” wording (rationale)



This nonpartisan campaign involves signing up people on Twitter **to make a post encouraging their friends and followers to sign a petition** opposing the movement to defund the police.

I have joined a campaign to oppose defunding the police: bit.ly/3DK3UEr. Before joining, I was shown this article written by a Princeton professor on the strong scientific evidence that defunding the police would increase violent crime: https://www.washingtonpost.com/outlook/2020/06/12/defund-police-violent-crime/

washingtonpost.com

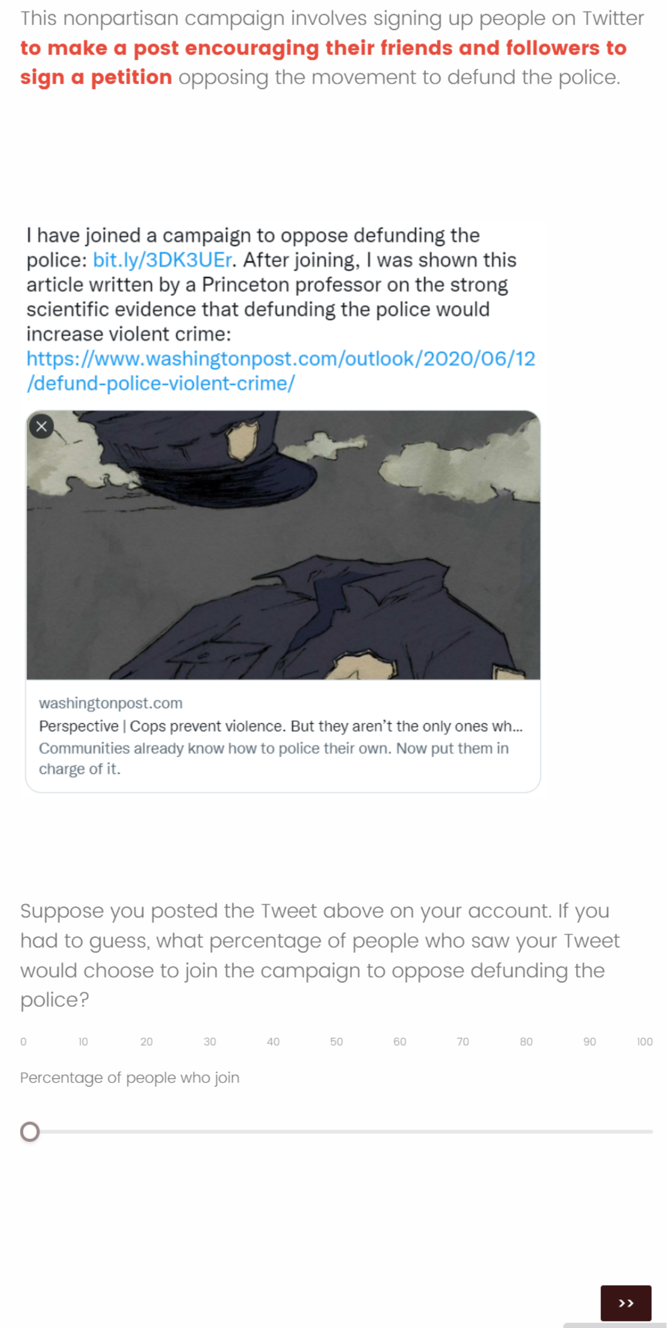
Perspective | Cops prevent violence. But they aren't the only ones wh...

Communities already know how to police their own. Now put them in charge of it.

Suppose you posted the Tweet above on your account. If you had to guess, what percentage of people who saw your Tweet would choose to join the campaign to oppose defunding the police?

Percentage of people who join

E.12.2 Treatment: "After" wording (no rationale)



This nonpartisan campaign involves signing up people on Twitter **to make a post encouraging their friends and followers to sign a petition** opposing the movement to defund the police.

I have joined a campaign to oppose defunding the police: bit.ly/3DK3UEr. After joining, I was shown this article written by a Princeton professor on the strong scientific evidence that defunding the police would increase violent crime: https://www.washingtonpost.com/outlook/2020/06/12/defund-police-violent-crime/

washingtonpost.com

Perspective | Cops prevent violence. But they aren't the only ones wh...

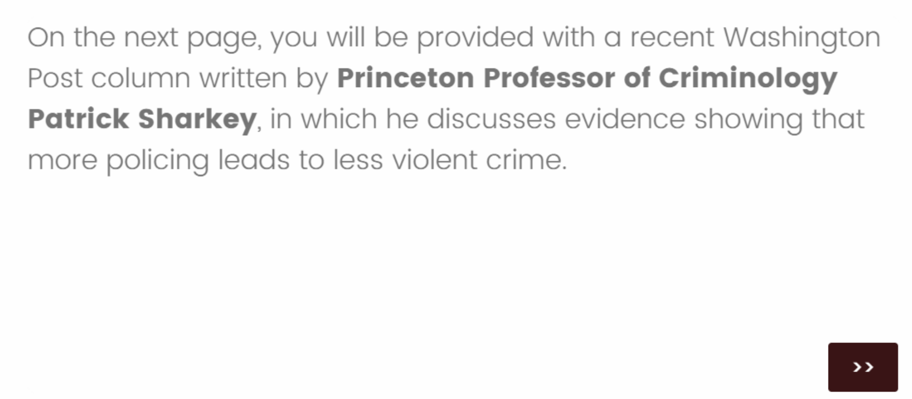
Communities already know how to police their own. Now put them in charge of it.

Suppose you posted the Tweet above on your account. If you had to guess, what percentage of people who saw your Tweet would choose to join the campaign to oppose defunding the police?

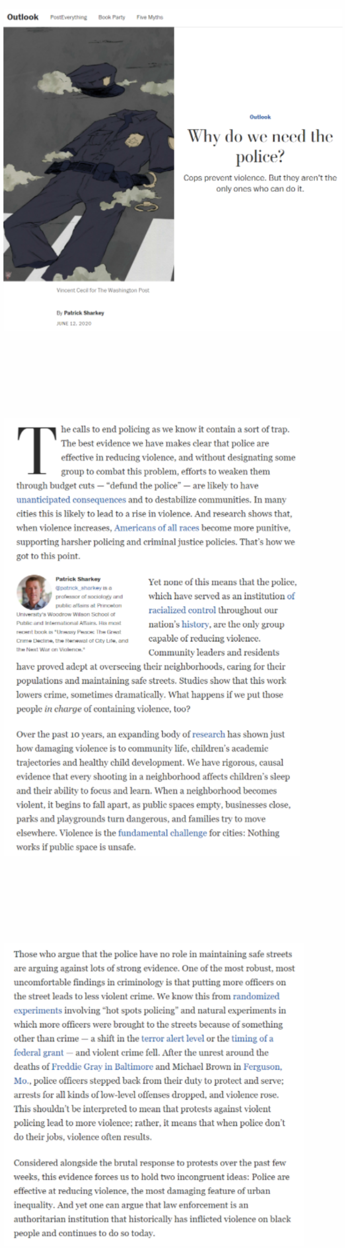
Percentage of people who join

E.13 Auxiliary Experiment 5: Open-ended explanations of preferred anti-defunding Tweet – Democrats

E.13.1 Pre-treatment questions



On the next page, you will be provided with a recent Washington Post column written by **Princeton Professor of Criminology Patrick Sharkey**, in which he discusses evidence showing that more policing leads to less violent crime.



Outlook

Why do we need the police?

Cops prevent violence. But they aren't the only ones who can do it.

Vincent Cecil for The Washington Post

By **Patrick Sharkey**

JUNE 12, 2020

The calls to end policing as we know it contain a sort of trap. The best evidence we have makes clear that police are effective in reducing violence, and without designating some group to combat this problem, efforts to weaken them through budget cuts — "defund the police" — are likely to have unanticipated consequences and to destabilize communities. In many cities this is likely to lead to a rise in violence. And research shows that, when violence increases, Americans of all races become more punitive, supporting harsher policing and criminal justice policies. That's how we got to this point.

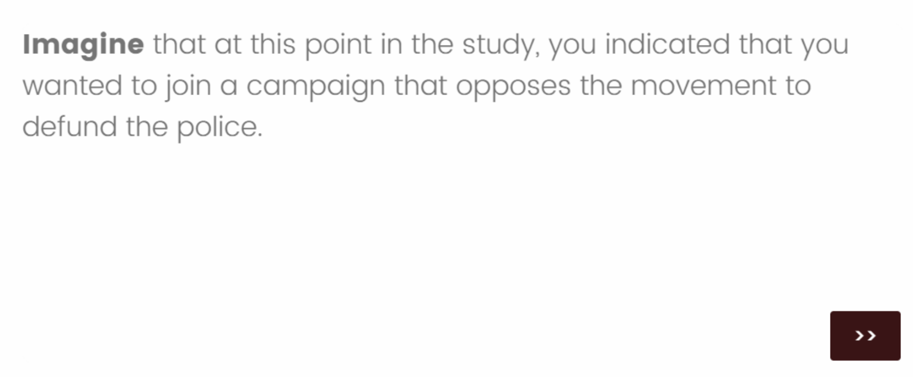
Patrick Sharkey @patrick_sharkey is a professor of sociology and public affairs at Princeton University's Woodrow Wilson School of Public and International Affairs. His most recent book is "Uneasy Peace: The Great Crime Decline, the Renewal of City Life, and the Next War on Violence."

Yet none of this means that the police, which have served as an institution of racialized control throughout our nation's history, are the only group capable of reducing violence. Community leaders and residents have proved adept at overseeing their neighborhoods, caring for their populations and maintaining safe streets. Studies show that this work lowers crime, sometimes dramatically. What happens if we put those people *in charge* of containing violence, too?

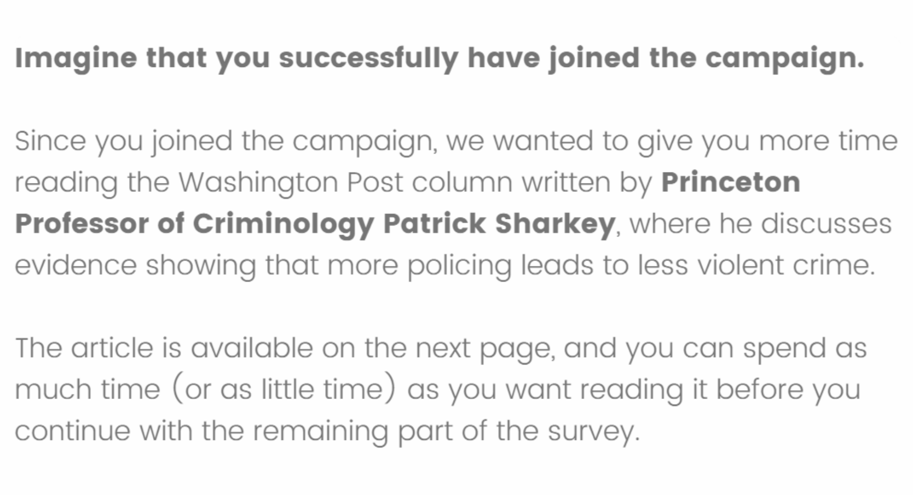
Over the past 10 years, an expanding body of research has shown just how damaging violence is to community life, children's academic trajectories and healthy child development. We have rigorous, causal evidence that every shooting in a neighborhood affects children's sleep and their ability to focus and learn. When a neighborhood becomes violent, it begins to fall apart, as public spaces empty, businesses close, parks and playgrounds turn dangerous, and families try to move elsewhere. Violence is the fundamental challenge for cities: Nothing works if public space is unsafe.

Those who argue that the police have no role in maintaining safe streets are arguing against lots of strong evidence. One of the most robust, most uncomfortable findings in criminology is that putting more officers on the street leads to less violent crime. We know this from randomized experiments involving "hot spots policing" and natural experiments in which more officers were brought to the streets because of something other than crime — a shift in the terror alert level or the timing of a federal grant — and violent crime fell. After the unrest around the deaths of Freddie Gray in Baltimore and Michael Brown in Ferguson, Mo., police officers stepped back from their duty to protect and serve; arrests for all kinds of low-level offenses dropped, and violence rose. This shouldn't be interpreted to mean that protests against violent policing lead to more violence; rather, it means that when police don't do their jobs, violence often results.

Considered alongside the brutal response to protests over the past few weeks, this evidence forces us to hold two incongruent ideas: Police are effective at reducing violence, the most damaging feature of urban inequality. And yet one can argue that law enforcement is an authoritarian institution that historically has inflicted violence on black people and continues to do so today.



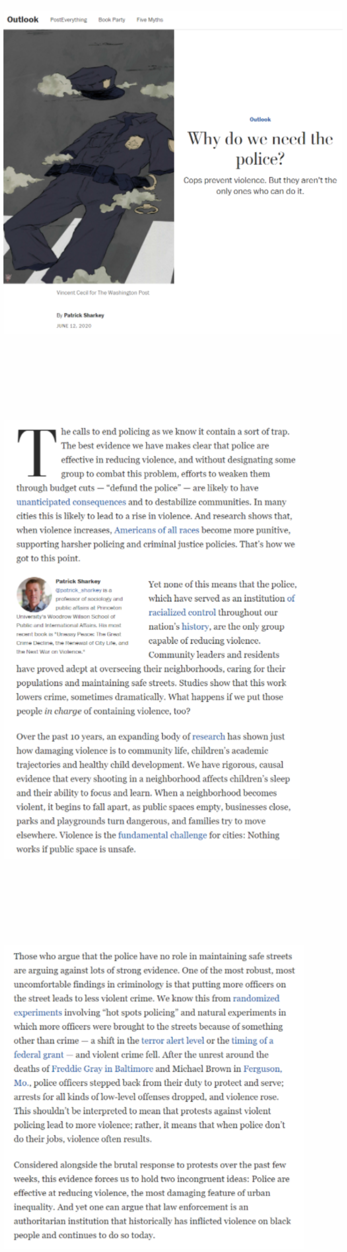
Imagine that at this point in the study, you indicated that you wanted to join a campaign that opposes the movement to defund the police.



Imagine that you successfully have joined the campaign.

Since you joined the campaign, we wanted to give you more time reading the Washington Post column written by **Princeton Professor of Criminology Patrick Sharkey**, where he discusses evidence showing that more policing leads to less violent crime.

The article is available on the next page, and you can spend as much time (or as little time) as you want reading it before you continue with the remaining part of the survey.



Outlook

Why do we need the police?

Cops prevent violence. But they aren't the only ones who can do it.

Vincent Cecil for The Washington Post

By **Patrick Sharkey**
JUNE 12, 2020

The calls to end policing as we know it contain a sort of trap. The best evidence we have makes clear that police are effective in reducing violence, and without designating some group to combat this problem, efforts to weaken them through budget cuts — "defund the police" — are likely to have unanticipated consequences and to destabilize communities. In many cities this is likely to lead to a rise in violence. And research shows that, when violence increases, Americans of all races become more punitive, supporting harsher policing and criminal justice policies. That's how we got to this point.

Yet none of this means that the police, which have served as an institution of racialized control throughout our nation's history, are the only group capable of reducing violence. Community leaders and residents have proved adept at overseeing their neighborhoods, caring for their populations and maintaining safe streets. Studies show that this work lowers crime, sometimes dramatically. What happens if we put those people *in charge* of containing violence, too?

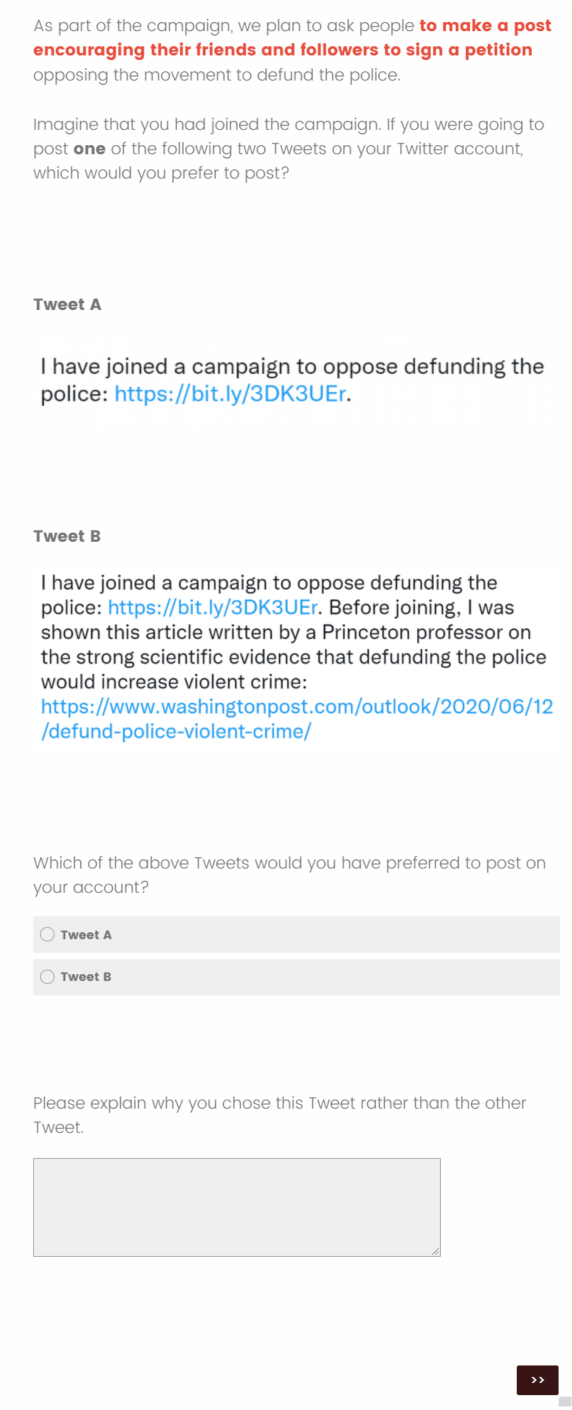
Patrick Sharkey
@patrick_sharkey is a professor of sociology and public affairs at Princeton University's Woodrow Wilson School of Public and International Affairs. His most recent book is "Uneasy Peace: The Great Crime Decline, the Renewal of City Life, and the Next War on Violence."

Over the past 10 years, an expanding body of research has shown just how damaging violence is to community life, children's academic trajectories and healthy child development. We have rigorous, causal evidence that every shooting in a neighborhood affects children's sleep and their ability to focus and learn. When a neighborhood becomes violent, it begins to fall apart, as public spaces empty, businesses close, parks and playgrounds turn dangerous, and families try to move elsewhere. Violence is the fundamental challenge for cities: Nothing works if public space is unsafe.

Those who argue that the police have no role in maintaining safe streets are arguing against lots of strong evidence. One of the most robust, most uncomfortable findings in criminology is that putting more officers on the street leads to less violent crime. We know this from randomized experiments involving "hot spots policing" and natural experiments in which more officers were brought to the streets because of something other than crime — a shift in the terror alert level or the timing of a federal grant — and violent crime fell. After the unrest around the deaths of Freddie Gray in Baltimore and Michael Brown in Ferguson, Mo., police officers stepped back from their duty to protect and serve; arrests for all kinds of low-level offenses dropped, and violence rose. This shouldn't be interpreted to mean that protests against violent policing lead to more violence; rather, it means that when police don't do their jobs, violence often results.

Considered alongside the brutal response to protests over the past few weeks, this evidence forces us to hold two incongruent ideas: Police are effective at reducing violence, the most damaging feature of urban inequality. And yet one can argue that law enforcement is an authoritarian institution that historically has inflicted violence on black people and continues to do so today.

E.13.2 Treatment: “Before” wording (rationale)



As part of the campaign, we plan to ask people **to make a post encouraging their friends and followers to sign a petition** opposing the movement to defund the police.

Imagine that you had joined the campaign. If you were going to post **one** of the following two Tweets on your Twitter account, which would you prefer to post?

Tweet A

I have joined a campaign to oppose defunding the police: https://bit.ly/3DK3UEr.

Tweet B

I have joined a campaign to oppose defunding the police: https://bit.ly/3DK3UEr. Before joining, I was shown this article written by a Princeton professor on the strong scientific evidence that defunding the police would increase violent crime: https://www.washingtonpost.com/outlook/2020/06/12/defund-police-violent-crime/

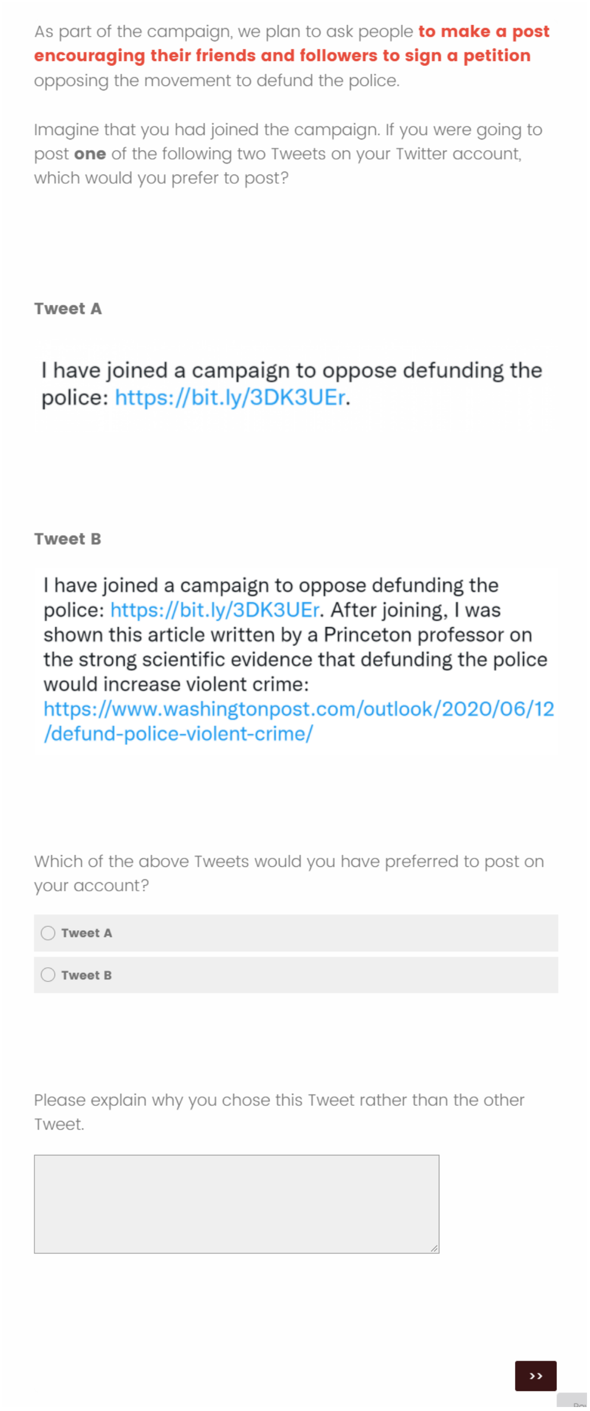
Which of the above Tweets would you have preferred to post on your account?

○ Tweet A

○ Tweet B

Please explain why you chose this Tweet rather than the other Tweet.

E.13.3 Treatment: "After" wording (no rationale)



As part of the campaign, we plan to ask people **to make a post encouraging their friends and followers to sign a petition** opposing the movement to defund the police.

Imagine that you had joined the campaign. If you were going to post **one** of the following two Tweets on your Twitter account, which would you prefer to post?

Tweet A

I have joined a campaign to oppose defunding the police: https://bit.ly/3DK3UEr.

Tweet B

I have joined a campaign to oppose defunding the police: https://bit.ly/3DK3UEr. After joining, I was shown this article written by a Princeton professor on the strong scientific evidence that defunding the police would increase violent crime: https://www.washingtonpost.com/outlook/2020/06/12/defund-police-violent-crime/

Which of the above Tweets would you have preferred to post on your account?

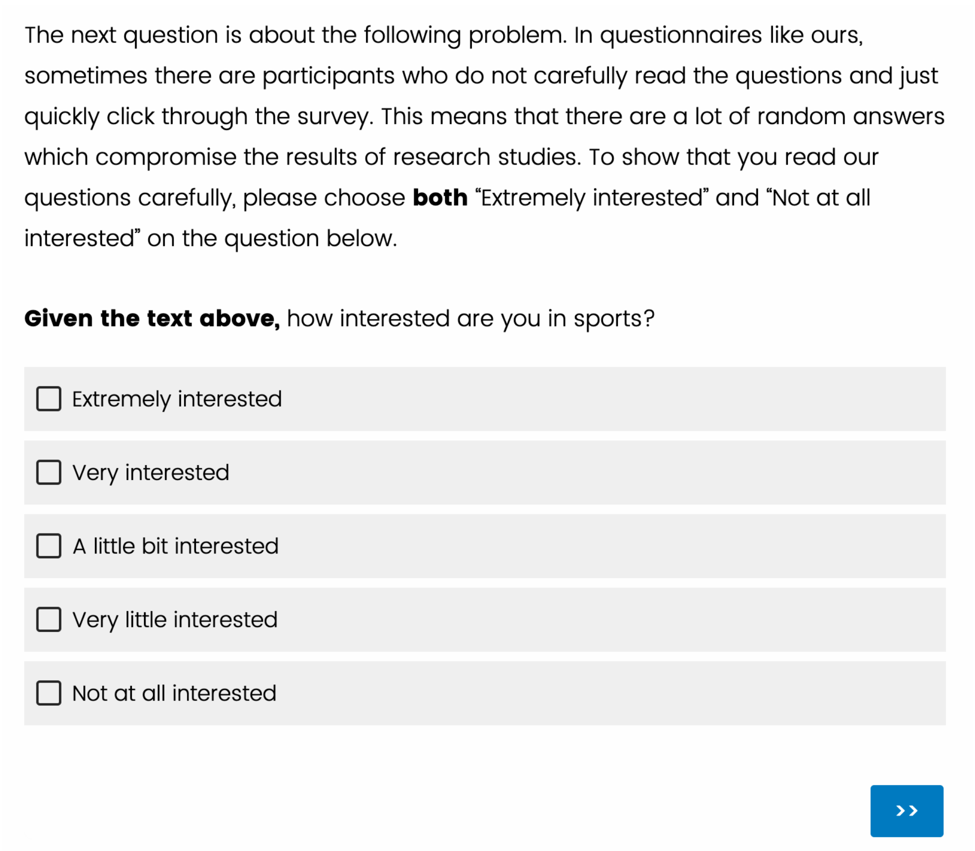
○ **Tweet A**

○ **Tweet B**

Please explain why you chose this Tweet rather than the other Tweet.

E.14 Auxiliary Experiment 6: Credibility and social cover – Democrats

E.14.1 Attention screener



The next question is about the following problem. In questionnaires like ours, sometimes there are participants who do not carefully read the questions and just quickly click through the survey. This means that there are a lot of random answers which compromise the results of research studies. To show that you read our questions carefully, please choose **both** "Extremely interested" and "Not at all interested" on the question below.

Given the text above, how interested are you in sports?

□ Extremely interested

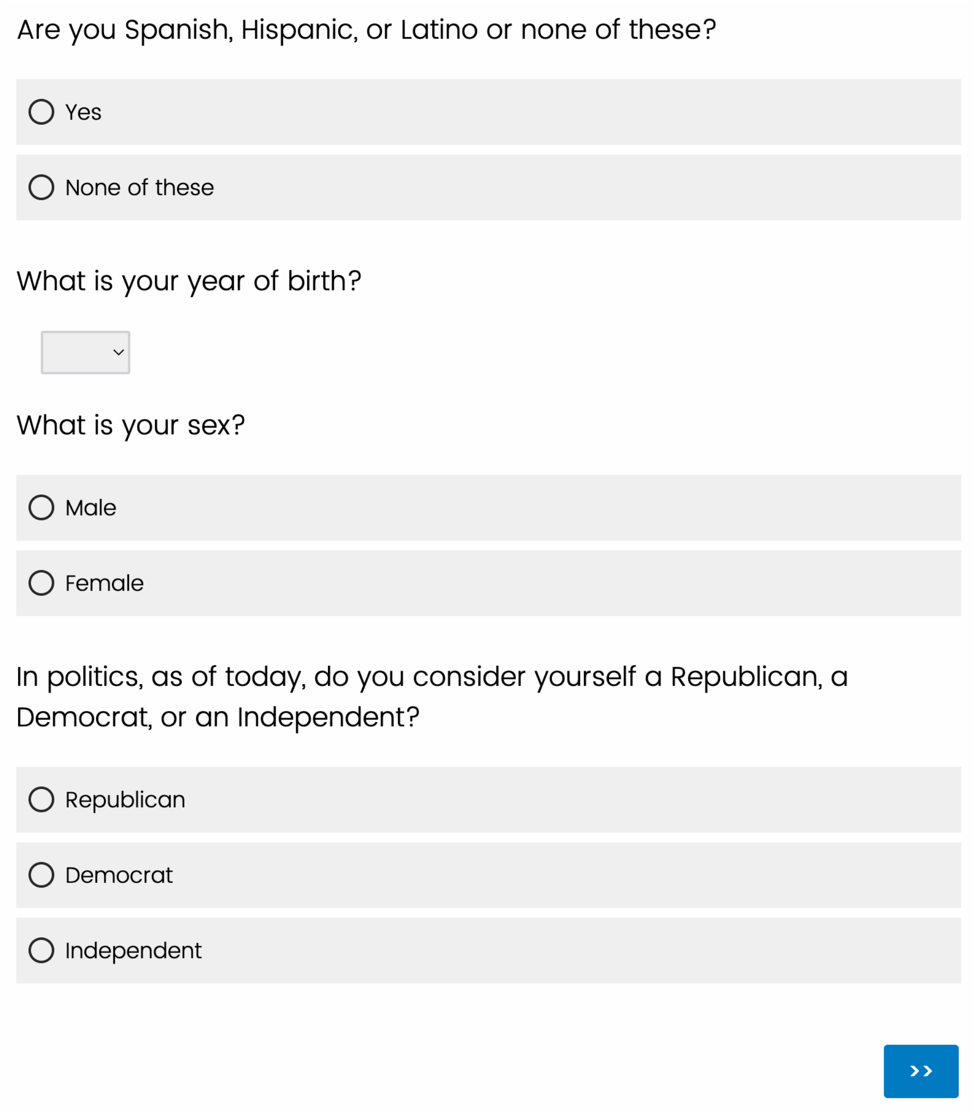
□ Very interested

□ A little bit interested

□ Very little interested

□ Not at all interested

E.14.2 Background questions



Are you Spanish, Hispanic, or Latino or none of these?

○ Yes

○ None of these

What is your year of birth?

What is your sex?

○ Male

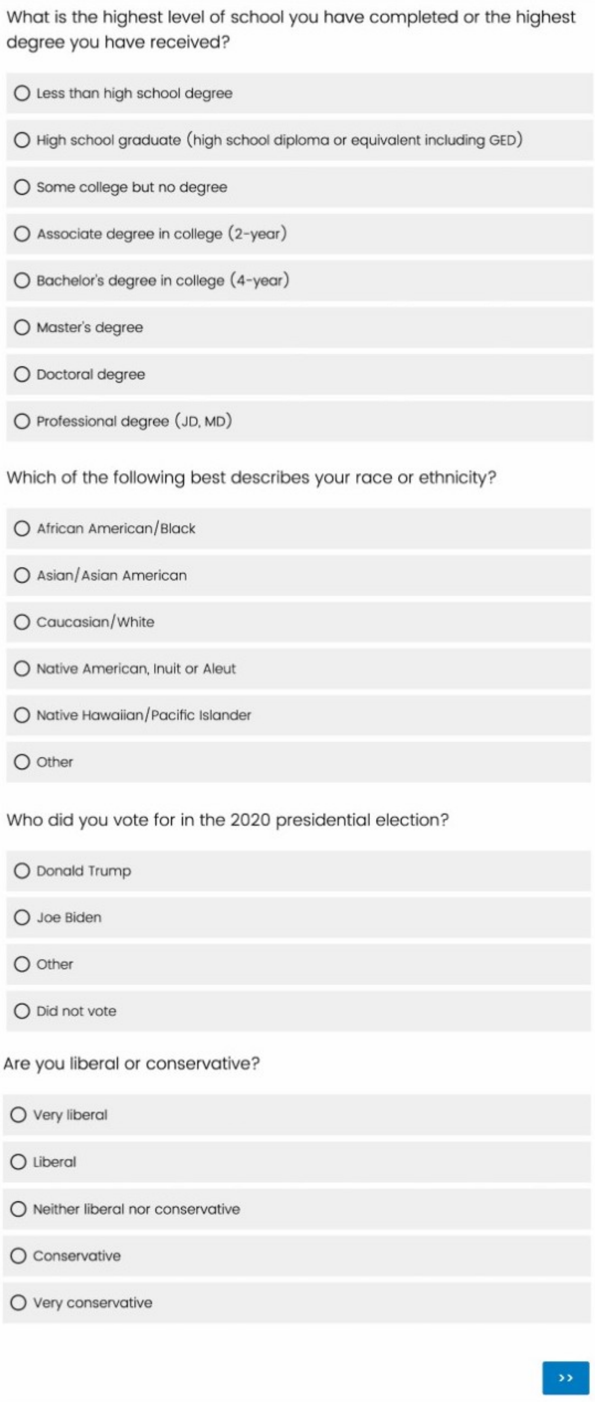
○ Female

In politics, as of today, do you consider yourself a Republican, a Democrat, or an Independent?

○ Republican

○ Democrat

○ Independent



What is the highest level of school you have completed or the highest degree you have received?

○ Less than high school degree

○ High school graduate (high school diploma or equivalent including GED)

○ Some college but no degree

○ Associate degree in college (2-year)

○ Bachelor's degree in college (4-year)

○ Master's degree

○ Doctoral degree

○ Professional degree (JD, MD)

Which of the following best describes your race or ethnicity?

○ African American/Black

○ Asian/Asian American

○ Caucasian/White

○ Native American, Inuit or Aleut

○ Native Hawaiian/Pacific Islander

○ Other

Who did you vote for in the 2020 presidential election?

○ Donald Trump

○ Joe Biden

○ Other

○ Did not vote

Are you liberal or conservative?

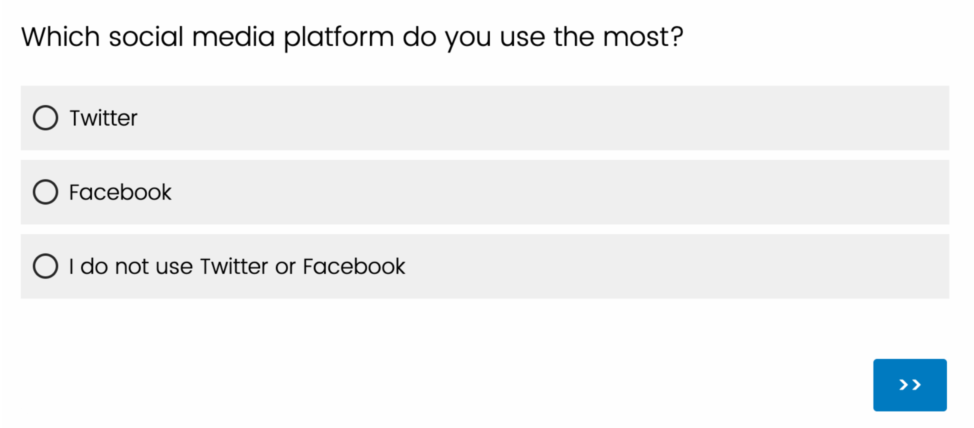
○ Very liberal

○ Liberal

○ Neither liberal nor conservative

○ Conservative

○ Very conservative



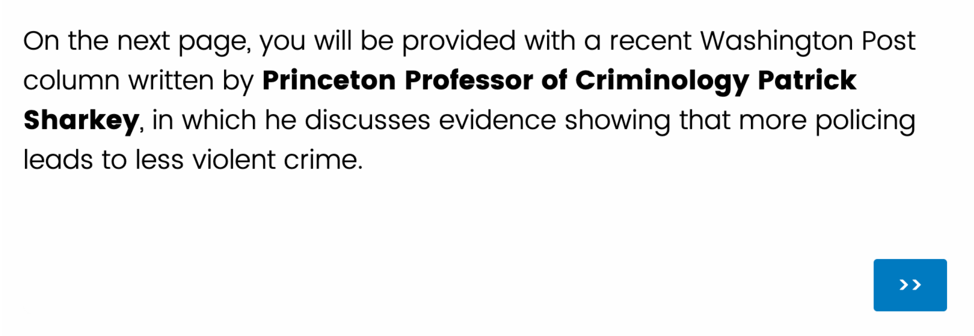
Which social media platform do you use the most?

○ Twitter

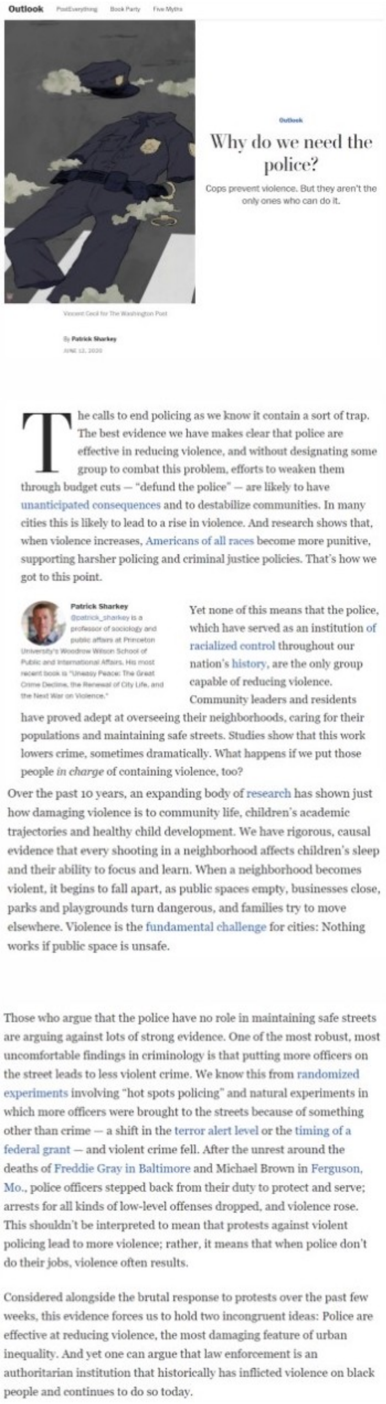
○ Facebook

○ I do not use Twitter or Facebook

E.14.3 Pre-treatment outcomes



On the next page, you will be provided with a recent Washington Post column written by **Princeton Professor of Criminology Patrick Sharkey**, in which he discusses evidence showing that more policing leads to less violent crime.



Outlook

Why do we need the police?

Cops prevent violence. But they aren't the only ones who can do it.

Vincent Cecil for The Washington Post

By Patrick Sharkey
JUNE 12, 2020

The calls to end policing as we know it contain a sort of trap. The best evidence we have makes clear that police are effective in reducing violence, and without designating some group to combat this problem, efforts to weaken them through budget cuts — "defund the police" — are likely to have unanticipated consequences and to destabilize communities. In many cities this is likely to lead to a rise in violence. And research shows that, when violence increases, Americans of all races become more punitive, supporting harsher policing and criminal justice policies. That's how we got to this point.

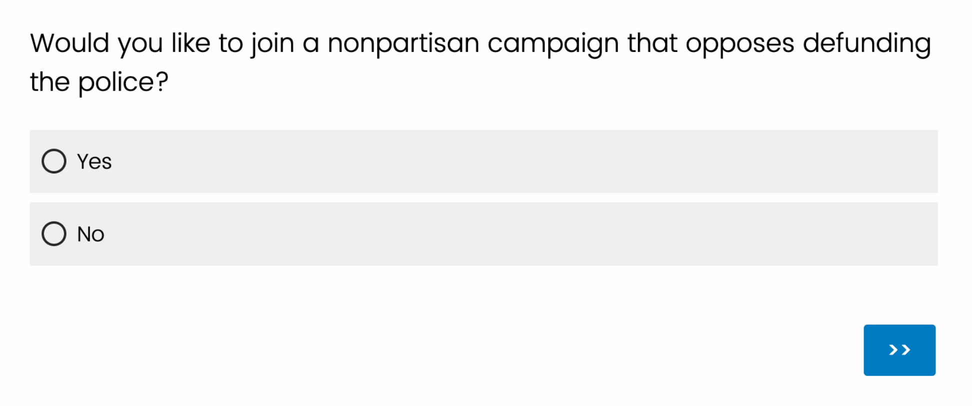
Patrick Sharkey
@patrick_sharkey is a professor of sociology and public affairs at Princeton University's Woodrow Wilson School of Public and International Affairs. His most recent book is "Uneasy Peace: The Great Crime Decline, the Renewal of City Life, and the Next War on Violence."

Yet none of this means that the police, which have served as an institution of racialized control throughout our nation's history, are the only group capable of reducing violence. Community leaders and residents have proved adept at overseeing their neighborhoods, caring for their populations and maintaining safe streets. Studies show that this work lowers crime, sometimes dramatically. What happens if we put those people *in charge* of containing violence, too?

Over the past 10 years, an expanding body of research has shown just how damaging violence is to community life, children's academic trajectories and healthy child development. We have rigorous, causal evidence that every shooting in a neighborhood affects children's sleep and their ability to focus and learn. When a neighborhood becomes violent, it begins to fall apart, as public spaces empty, businesses close, parks and playgrounds turn dangerous, and families try to move elsewhere. Violence is the fundamental challenge for cities: Nothing works if public space is unsafe.

Those who argue that the police have no role in maintaining safe streets are arguing against lots of strong evidence. One of the most robust, most uncomfortable findings in criminology is that putting more officers on the street leads to less violent crime. We know this from randomized experiments involving "hot spots policing" and natural experiments in which more officers were brought to the streets because of something other than crime — a shift in the terror alert level or the timing of a federal grant — and violent crime fell. After the unrest around the deaths of Freddie Gray in Baltimore and Michael Brown in Ferguson, Mo., police officers stepped back from their duty to protect and serve; arrests for all kinds of low-level offenses dropped, and violence rose. This shouldn't be interpreted to mean that protests against violent policing lead to more violence; rather, it means that when police don't do their jobs, violence often results.

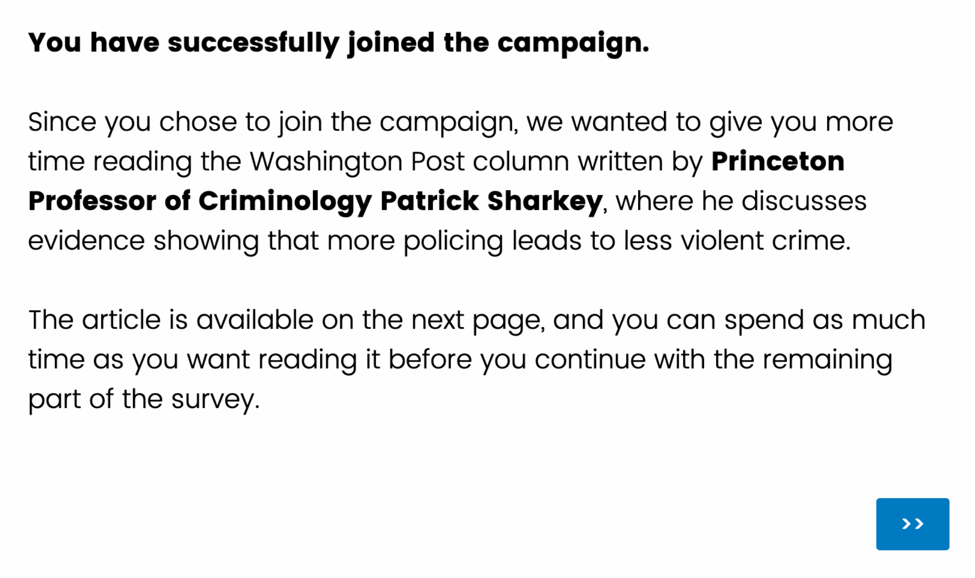
Considered alongside the brutal response to protests over the past few weeks, this evidence forces us to hold two incongruent ideas: Police are effective at reducing violence, the most damaging feature of urban inequality. And yet one can argue that law enforcement is an authoritarian institution that historically has inflicted violence on black people and continues to do so today.



Would you like to join a nonpartisan campaign that opposes defunding the police?

○ Yes

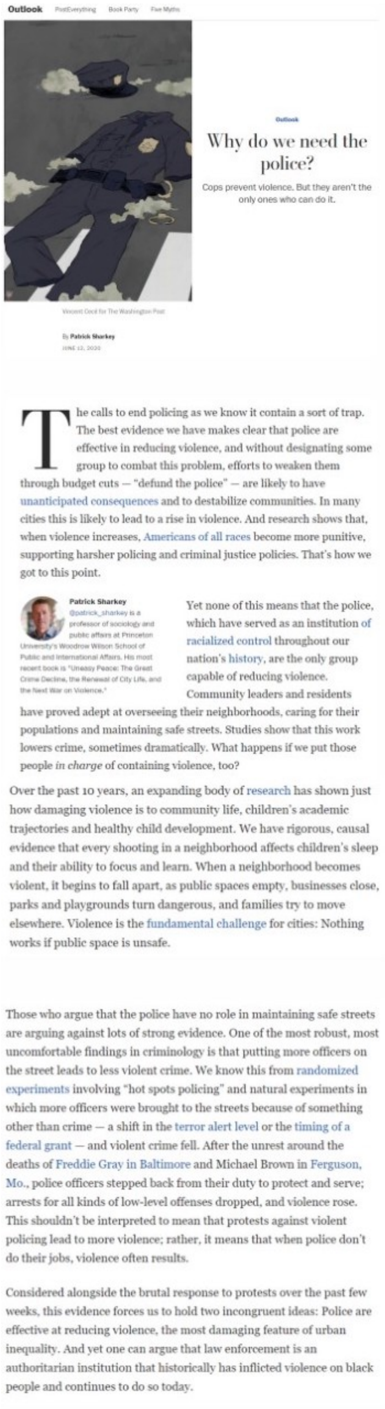
○ No



You have successfully joined the campaign.

Since you chose to join the campaign, we wanted to give you more time reading the Washington Post column written by **Princeton Professor of Criminology Patrick Sharkey**, where he discusses evidence showing that more policing leads to less violent crime.

The article is available on the next page, and you can spend as much time as you want reading it before you continue with the remaining part of the survey.



Why do we need the police?

Cops prevent violence. But they aren't the only ones who can do it.

Vincent Cecil for The Washington Post

By Patrick Sharkey

JUNE 12, 2020

The calls to end policing as we know it contain a sort of trap. The best evidence we have makes clear that police are effective in reducing violence, and without designating some group to combat this problem, efforts to weaken them through budget cuts — "defund the police" — are likely to have unanticipated consequences and to destabilize communities. In many cities this is likely to lead to a rise in violence. And research shows that, when violence increases, Americans of all races become more punitive, supporting harsher policing and criminal justice policies. That's how we got to this point.

Patrick Sharkey @patrick_sharkey is a professor of sociology and public affairs at Princeton University's Woodrow Wilson School of Public and International Affairs. His most recent book is "Uneasy Peace: The Great Crime Decline, the Renewal of City Life, and the Next War on Violence."

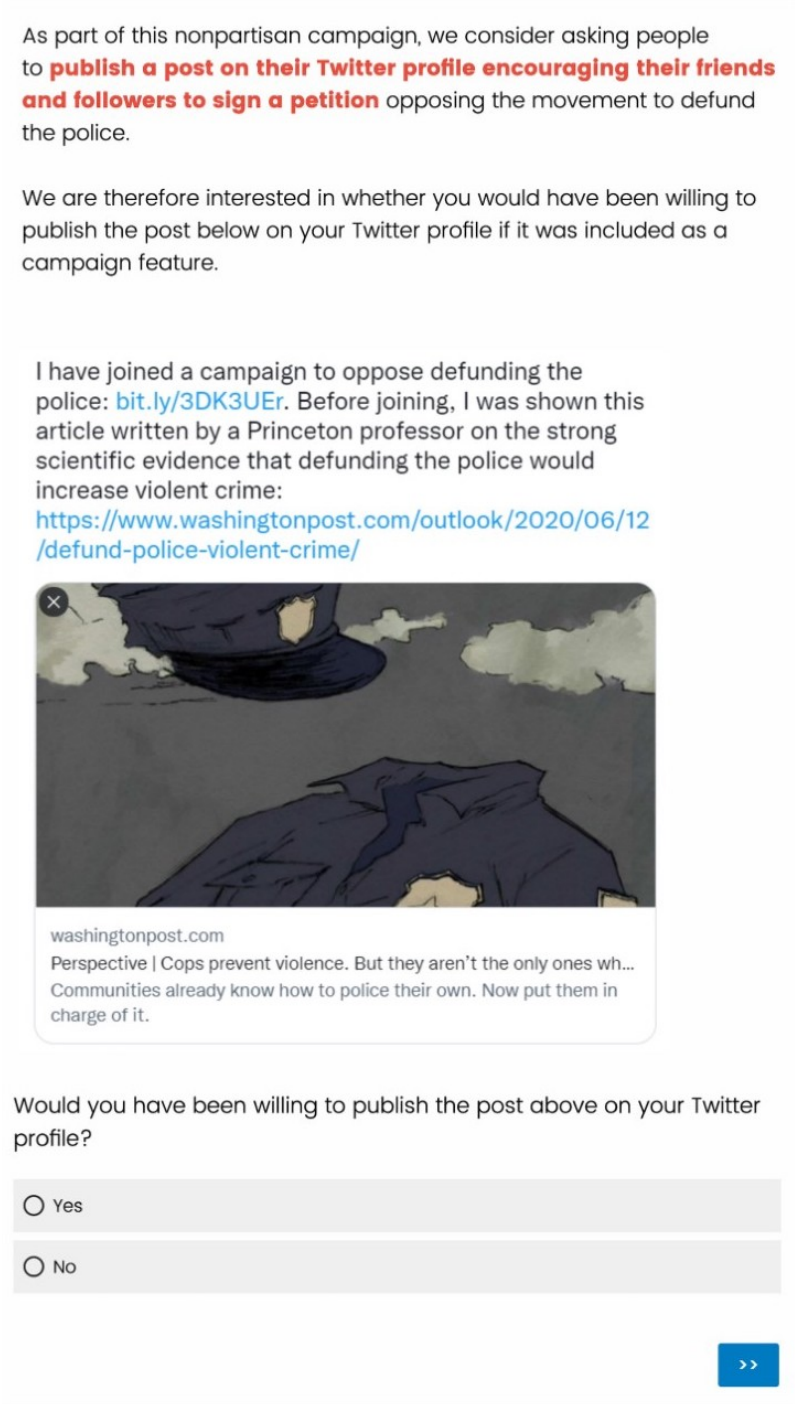
Yet none of this means that the police, which have served as an institution of racialized control throughout our nation's history, are the only group capable of reducing violence. Community leaders and residents have proved adept at overseeing their neighborhoods, caring for their populations and maintaining safe streets. Studies show that this work lowers crime, sometimes dramatically. What happens if we put those people *in charge* of containing violence, too?

Over the past 10 years, an expanding body of research has shown just how damaging violence is to community life, children's academic trajectories and healthy child development. We have rigorous, causal evidence that every shooting in a neighborhood affects children's sleep and their ability to focus and learn. When a neighborhood becomes violent, it begins to fall apart, as public spaces empty, businesses close, parks and playgrounds turn dangerous, and families try to move elsewhere. Violence is the fundamental challenge for cities: Nothing works if public space is unsafe.

Those who argue that the police have no role in maintaining safe streets are arguing against lots of strong evidence. One of the most robust, most uncomfortable findings in criminology is that putting more officers on the street leads to less violent crime. We know this from randomized experiments involving "hot spots policing" and natural experiments in which more officers were brought to the streets because of something other than crime — a shift in the terror alert level or the timing of a federal grant — and violent crime fell. After the unrest around the deaths of Freddie Gray in Baltimore and Michael Brown in Ferguson, Mo., police officers stepped back from their duty to protect and serve; arrests for all kinds of low-level offenses dropped, and violence rose. This shouldn't be interpreted to mean that protests against violent policing lead to more violence; rather, it means that when police don't do their jobs, violence often results.

Considered alongside the brutal response to protests over the past few weeks, this evidence forces us to hold two incongruent ideas: Police are effective at reducing violence, the most damaging feature of urban inequality. And yet one can argue that law enforcement is an authoritarian institution that historically has inflicted violence on black people and continues to do so today.

E.14.4 Treatment (higher-credibility): “Before” wording (rationale)



As part of this nonpartisan campaign, we consider asking people to **publish a post on their Twitter profile encouraging their friends and followers to sign a petition** opposing the movement to defund the police.

We are therefore interested in whether you would have been willing to publish the post below on your Twitter profile if it was included as a campaign feature.

I have joined a campaign to oppose defunding the police: bit.ly/3DK3UEr. Before joining, I was shown this article written by a Princeton professor on the strong scientific evidence that defunding the police would increase violent crime: https://www.washingtonpost.com/outlook/2020/06/12/defund-police-violent-crime/

washingtonpost.com

Perspective | Cops prevent violence. But they aren’t the only ones wh...

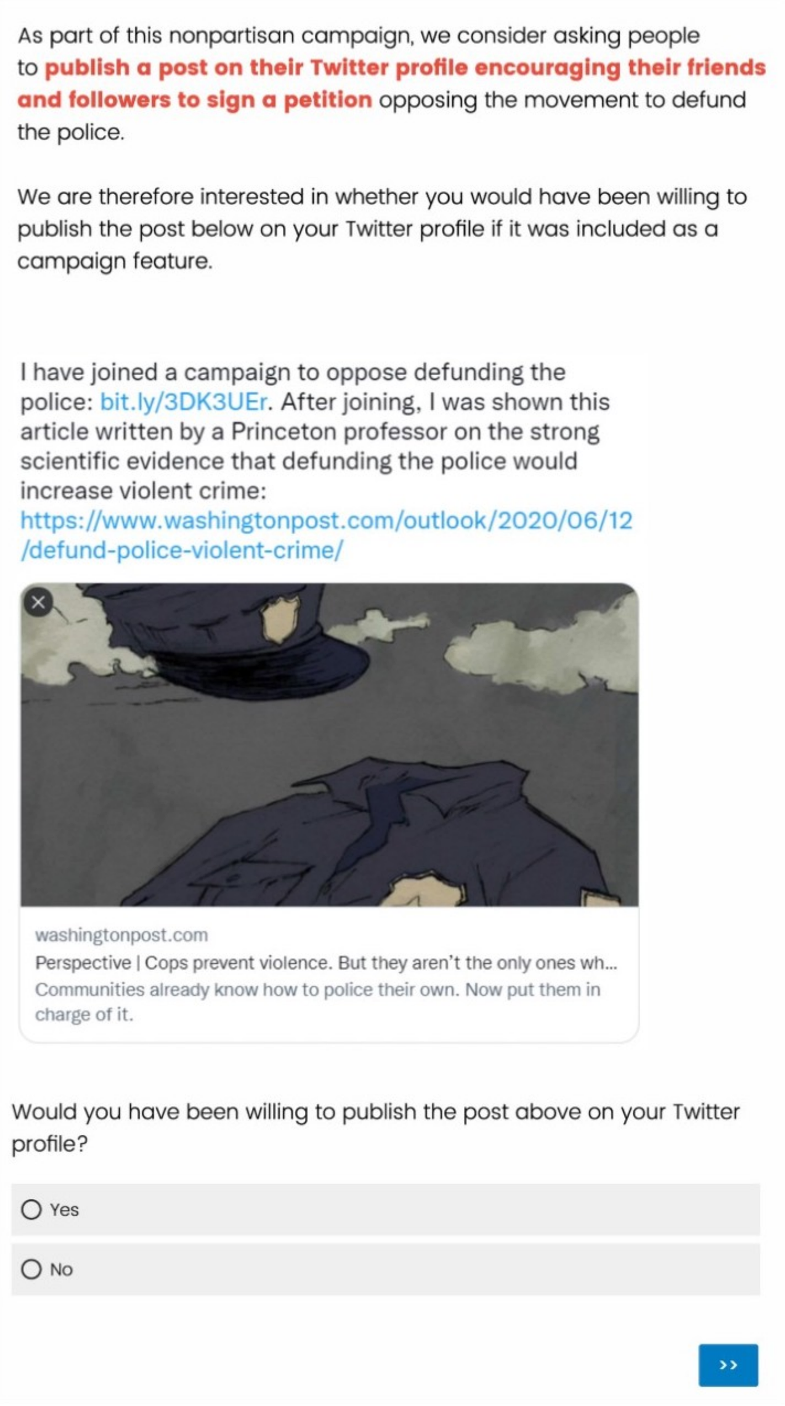
Communities already know how to police their own. Now put them in charge of it.

Would you have been willing to publish the post above on your Twitter profile?

○ Yes

○ No

E.14.5 Treatment (higher-credibility): “After” wording (no rationale)



As part of this nonpartisan campaign, we consider asking people to **publish a post on their Twitter profile encouraging their friends and followers to sign a petition** opposing the movement to defund the police.

We are therefore interested in whether you would have been willing to publish the post below on your Twitter profile if it was included as a campaign feature.

I have joined a campaign to oppose defunding the police: bit.ly/3DK3UEr. After joining, I was shown this article written by a Princeton professor on the strong scientific evidence that defunding the police would increase violent crime: https://www.washingtonpost.com/outlook/2020/06/12/defund-police-violent-crime/

washingtonpost.com

Perspective | Cops prevent violence. But they aren’t the only ones wh...

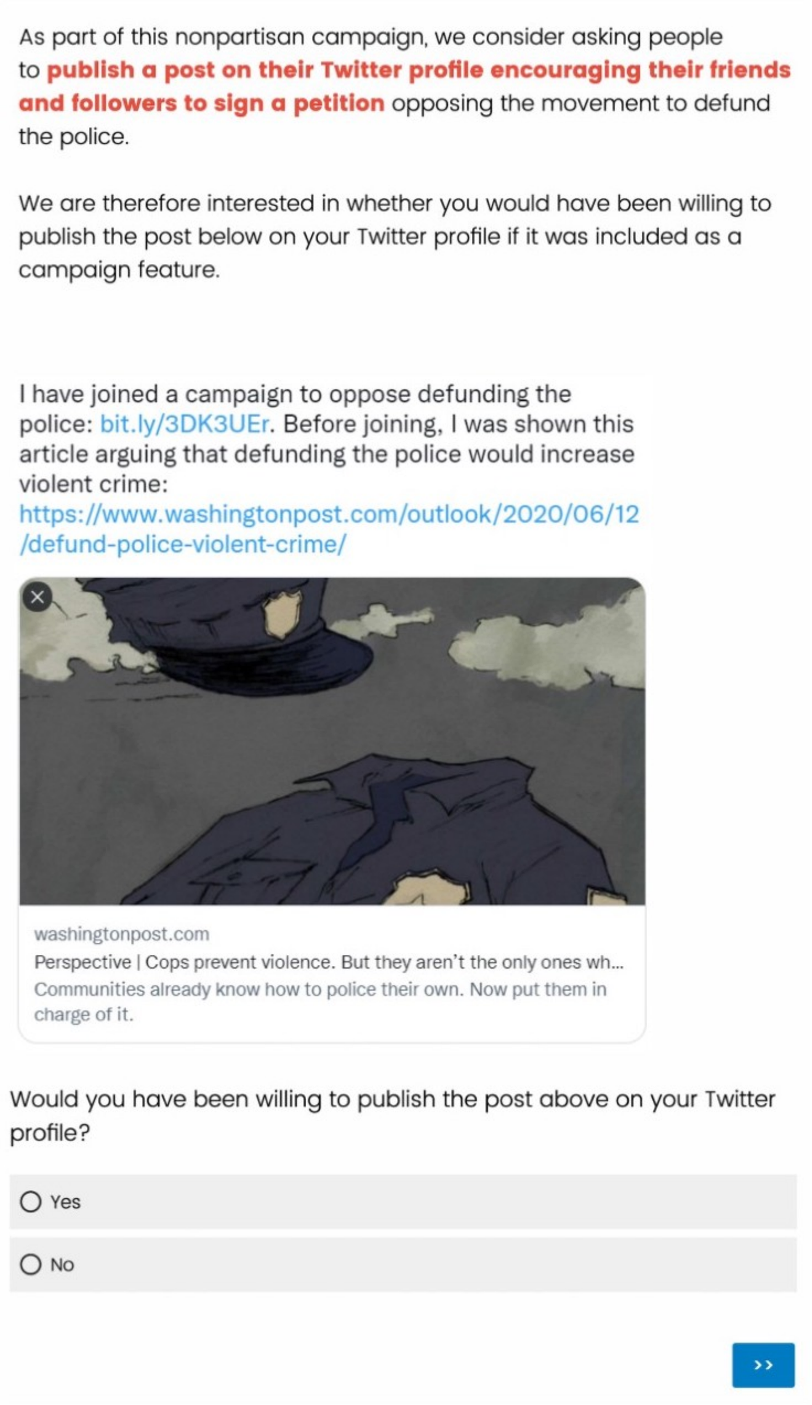
Communities already know how to police their own. Now put them in charge of it.

Would you have been willing to publish the post above on your Twitter profile?

○ Yes

○ No

E.14.6 Treatment (lower-credibility): "Before" wording (rationale)



As part of this nonpartisan campaign, we consider asking people to **publish a post on their Twitter profile encouraging their friends and followers to sign a petition** opposing the movement to defund the police.

We are therefore interested in whether you would have been willing to publish the post below on your Twitter profile if it was included as a campaign feature.

I have joined a campaign to oppose defunding the police: bit.ly/3DK3UEr. Before joining, I was shown this article arguing that defunding the police would increase violent crime: https://www.washingtonpost.com/outlook/2020/06/12/defund-police-violent-crime/

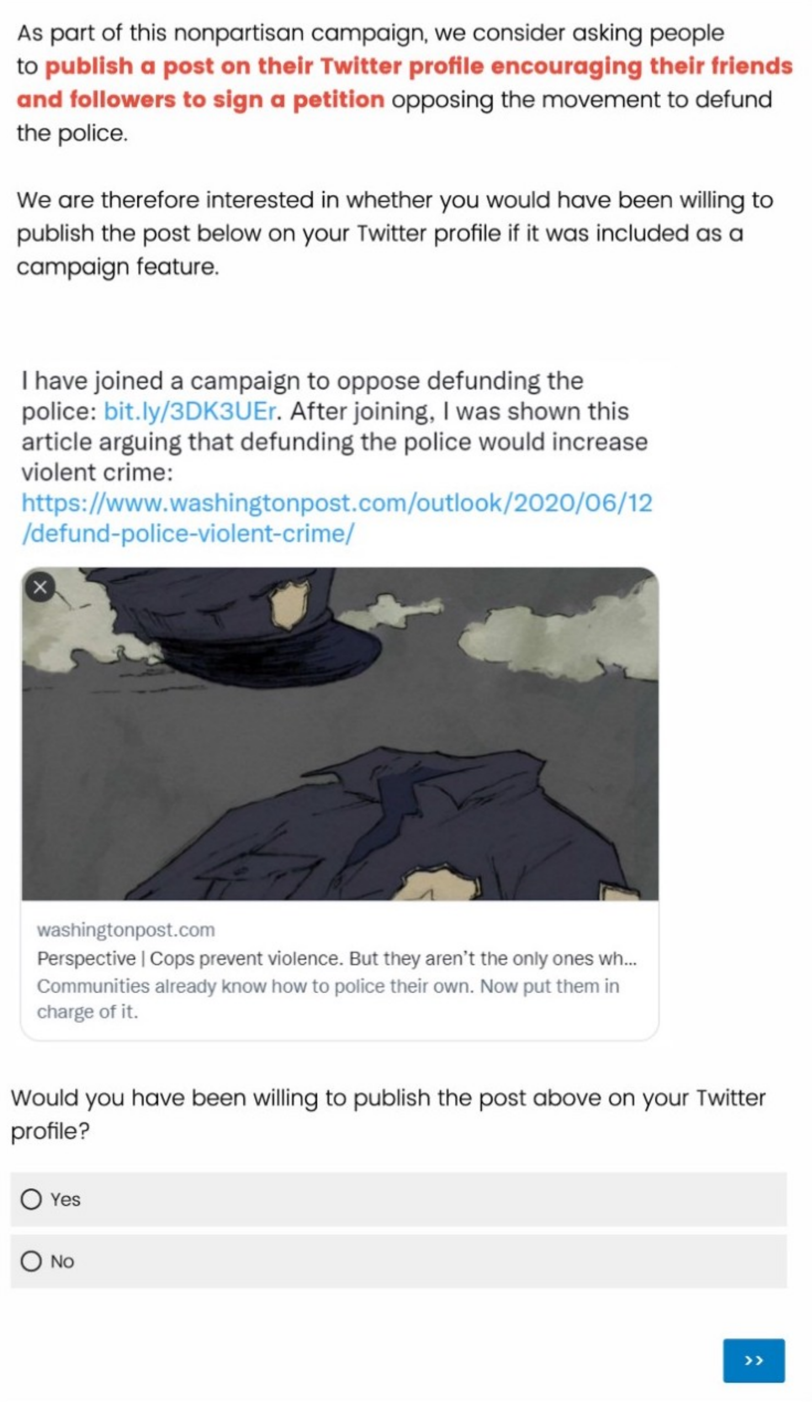
washingtonpost.com
Perspective | Cops prevent violence. But they aren't the only ones wh...
Communities already know how to police their own. Now put them in charge of it.

Would you have been willing to publish the post above on your Twitter profile?

○ Yes

○ No

E.14.7 Treatment (lower-credibility): "After" wording (no rationale)



As part of this nonpartisan campaign, we consider asking people to **publish a post on their Twitter profile encouraging their friends and followers to sign a petition** opposing the movement to defund the police.

We are therefore interested in whether you would have been willing to publish the post below on your Twitter profile if it was included as a campaign feature.

I have joined a campaign to oppose defunding the police: bit.ly/3DK3UEr. After joining, I was shown this article arguing that defunding the police would increase violent crime: https://www.washingtonpost.com/outlook/2020/06/12/defund-police-violent-crime/

washingtonpost.com

Perspective | Cops prevent violence. But they aren't the only ones wh...

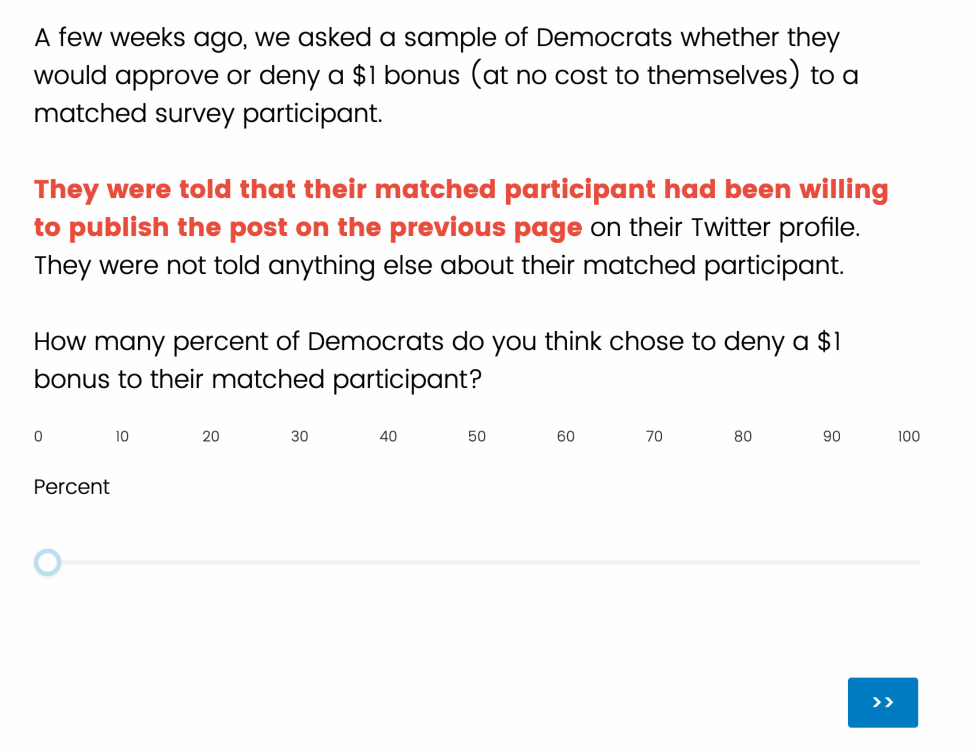
Communities already know how to police their own. Now put them in charge of it.

Would you have been willing to publish the post above on your Twitter profile?

○ Yes

○ No

E.14.8 Perceived social punishment



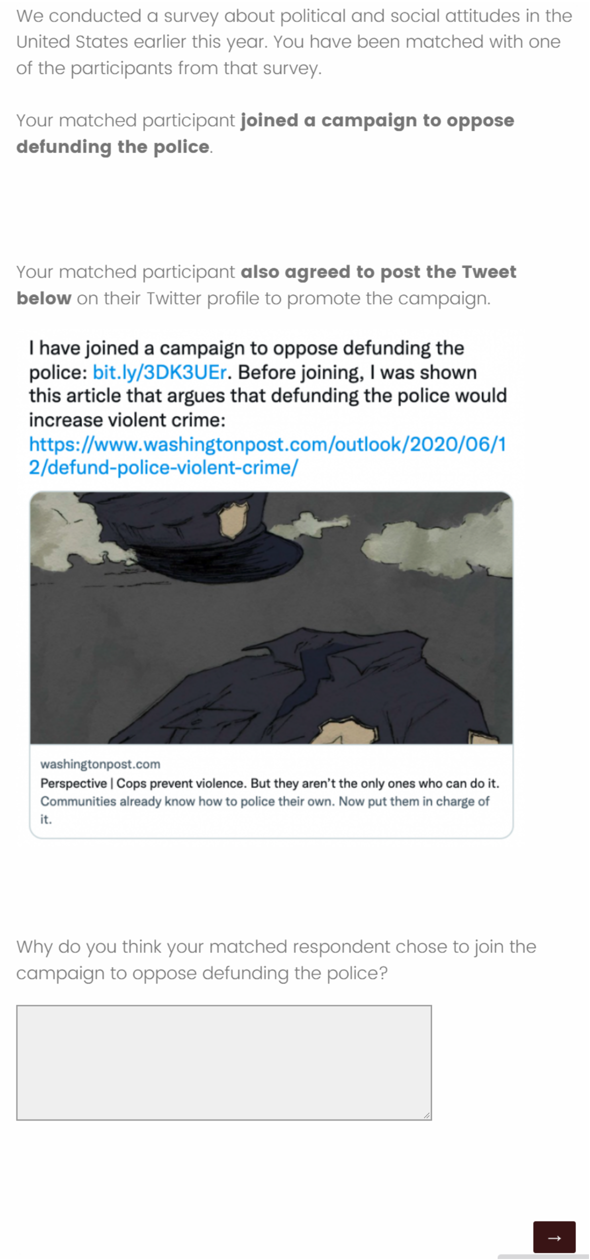
A few weeks ago, we asked a sample of Democrats whether they would approve or deny a $1 bonus (at no cost to themselves) to a matched survey participant.

They were told that their matched participant had been willing to publish the post on the previous page on their Twitter profile. They were not told anything else about their matched participant.

How many percent of Democrats do you think chose to deny a $1 bonus to their matched participant?

E.15 Auxiliary Experiments 7: Interpretation of dissent with low-credibility rationale – Democrats

E.15.1 Treatment: "Before" condition (rationale)



We conducted a survey about political and social attitudes in the United States earlier this year. You have been matched with one of the participants from that survey.

Your matched participant **joined a campaign to oppose defunding the police**.

Your matched participant **also agreed to post the Tweet below** on their Twitter profile to promote the campaign.

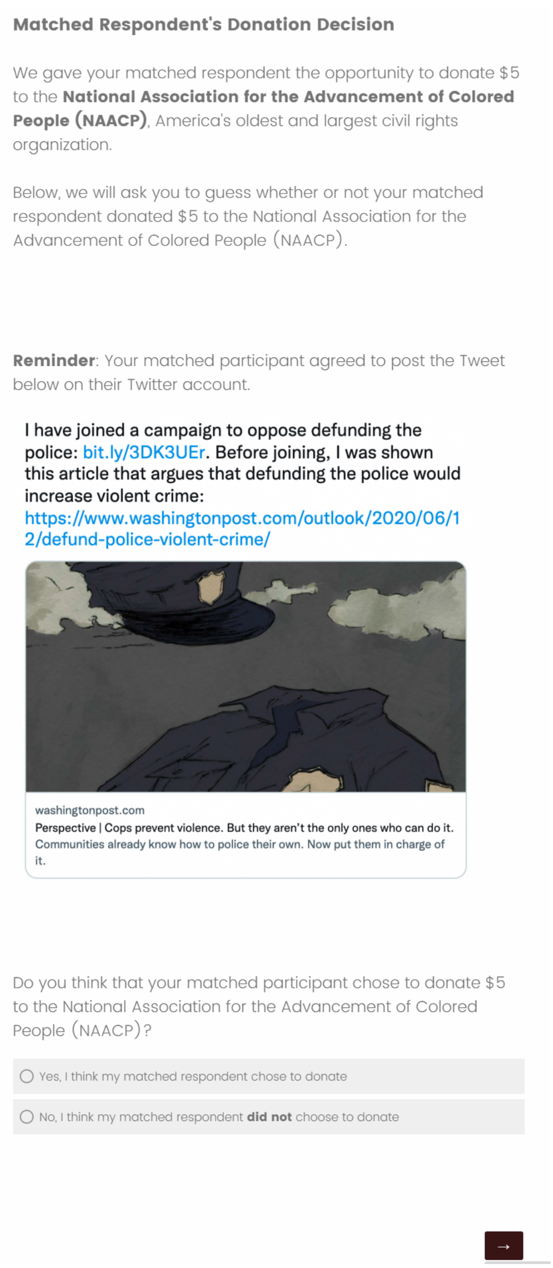
I have joined a campaign to oppose defunding the police: bit.ly/3DK3UEr. Before joining, I was shown this article that argues that defunding the police would increase violent crime: https://www.washingtonpost.com/outlook/2020/06/12/defund-police-violent-crime/

washingtonpost.com

Perspective | Cops prevent violence. But they aren't the only ones who can do it.

Communities already know how to police their own. Now put them in charge of it.

Why do you think your matched respondent chose to join the campaign to oppose defunding the police?



Matched Respondent's Donation Decision

We gave your matched respondent the opportunity to donate $5 to the **National Association for the Advancement of Colored People (NAACP)**, America's oldest and largest civil rights organization.

Below, we will ask you to guess whether or not your matched respondent donated $5 to the National Association for the Advancement of Colored People (NAACP).

Reminder: Your matched participant agreed to post the Tweet below on their Twitter account.

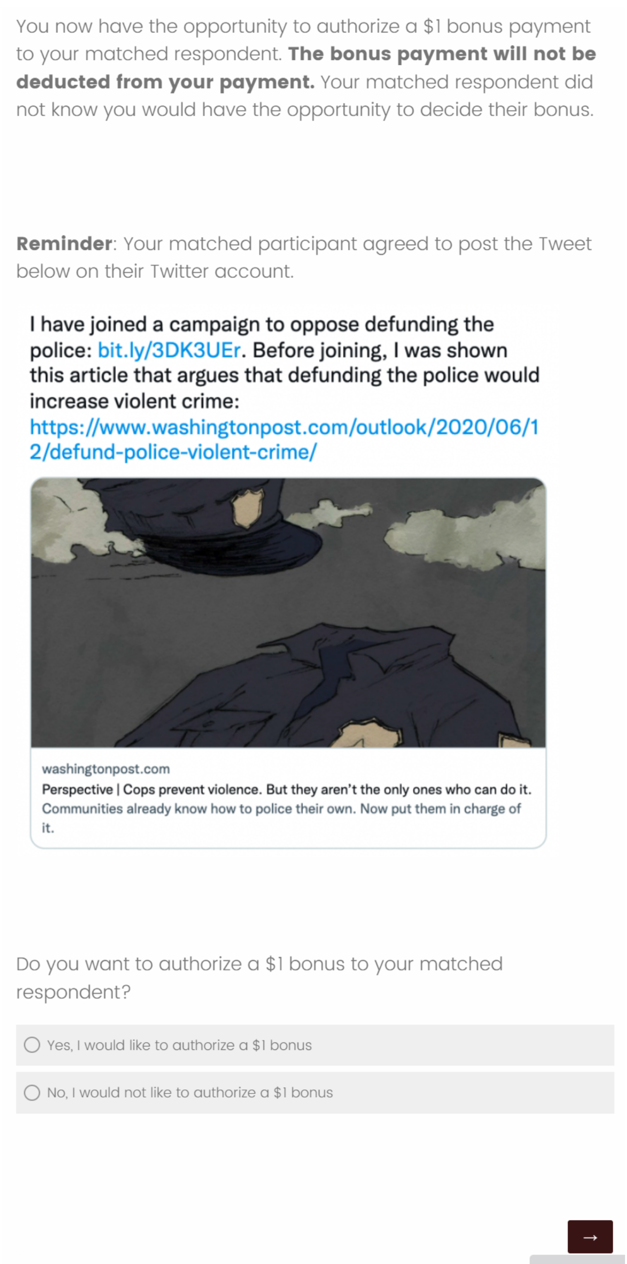
I have joined a campaign to oppose defunding the police: bit.ly/3DK3UEr. Before joining, I was shown this article that argues that defunding the police would increase violent crime: https://www.washingtonpost.com/outlook/2020/06/12/defund-police-violent-crime/

washingtonpost.com
Perspective | Cops prevent violence. But they aren't the only ones who can do it.
Communities already know how to police their own. Now put them in charge of it.

Do you think that your matched participant chose to donate $5 to the National Association for the Advancement of Colored People (NAACP)?

○ Yes, I think my matched respondent chose to donate

○ No, I think my matched respondent **did not** choose to donate



You now have the opportunity to authorize a $1 bonus payment to your matched respondent. **The bonus payment will not be deducted from your payment.** Your matched respondent did not know you would have the opportunity to decide their bonus.

Reminder: Your matched participant agreed to post the Tweet below on their Twitter account.

I have joined a campaign to oppose defunding the police: bit.ly/3DK3UEr. Before joining, I was shown this article that argues that defunding the police would increase violent crime: https://www.washingtonpost.com/outlook/2020/06/12/defund-police-violent-crime/

washingtonpost.com

Perspective | Cops prevent violence. But they aren't the only ones who can do it.

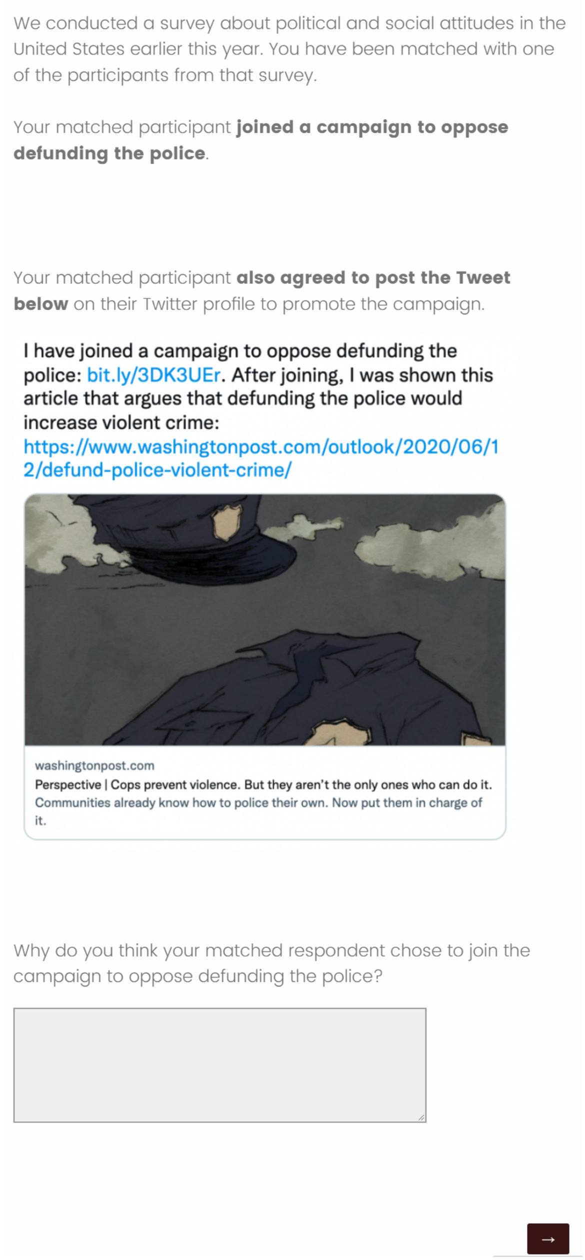
Communities already know how to police their own. Now put them in charge of it.

Do you want to authorize a $1 bonus to your matched respondent?

○ Yes, I would like to authorize a $1 bonus

○ No, I would not like to authorize a $1 bonus

E.15.2 Treatment: "After" condition (no rationale)



We conducted a survey about political and social attitudes in the United States earlier this year. You have been matched with one of the participants from that survey.

Your matched participant **joined a campaign to oppose defunding the police**.

Your matched participant **also agreed to post the Tweet below** on their Twitter profile to promote the campaign.

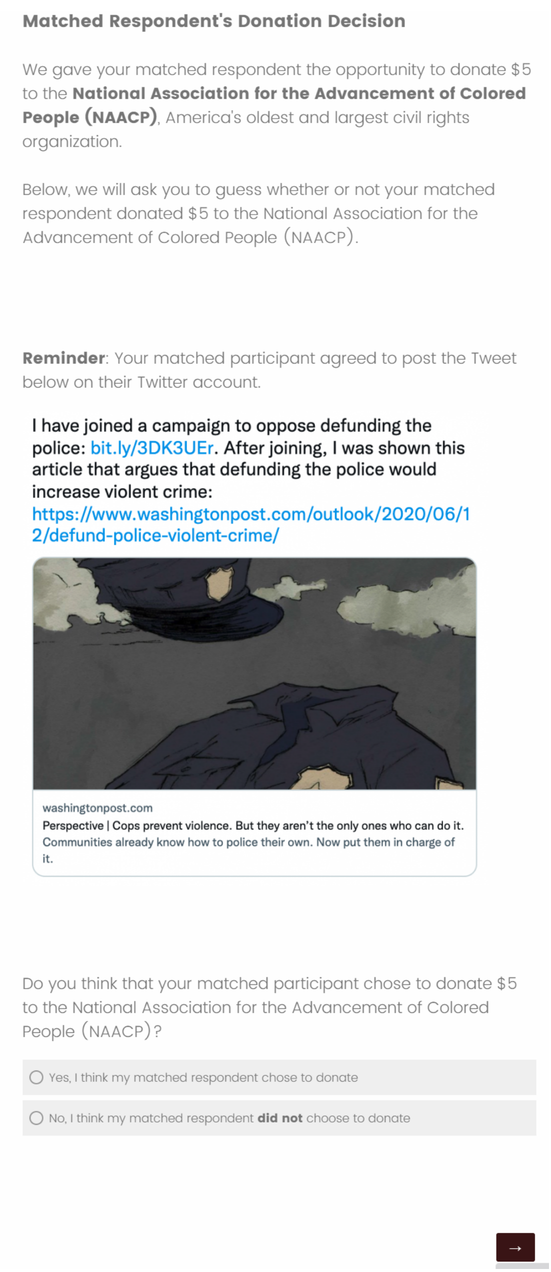
I have joined a campaign to oppose defunding the police: bit.ly/3DK3UEr. After joining, I was shown this article that argues that defunding the police would increase violent crime: https://www.washingtonpost.com/outlook/2020/06/12/defund-police-violent-crime/

washingtonpost.com

Perspective | Cops prevent violence. But they aren't the only ones who can do it.

Communities already know how to police their own. Now put them in charge of it.

Why do you think your matched respondent chose to join the campaign to oppose defunding the police?



Matched Respondent's Donation Decision

We gave your matched respondent the opportunity to donate $5 to the **National Association for the Advancement of Colored People (NAACP)**, America's oldest and largest civil rights organization.

Below, we will ask you to guess whether or not your matched respondent donated $5 to the National Association for the Advancement of Colored People (NAACP).

Reminder: Your matched participant agreed to post the Tweet below on their Twitter account.

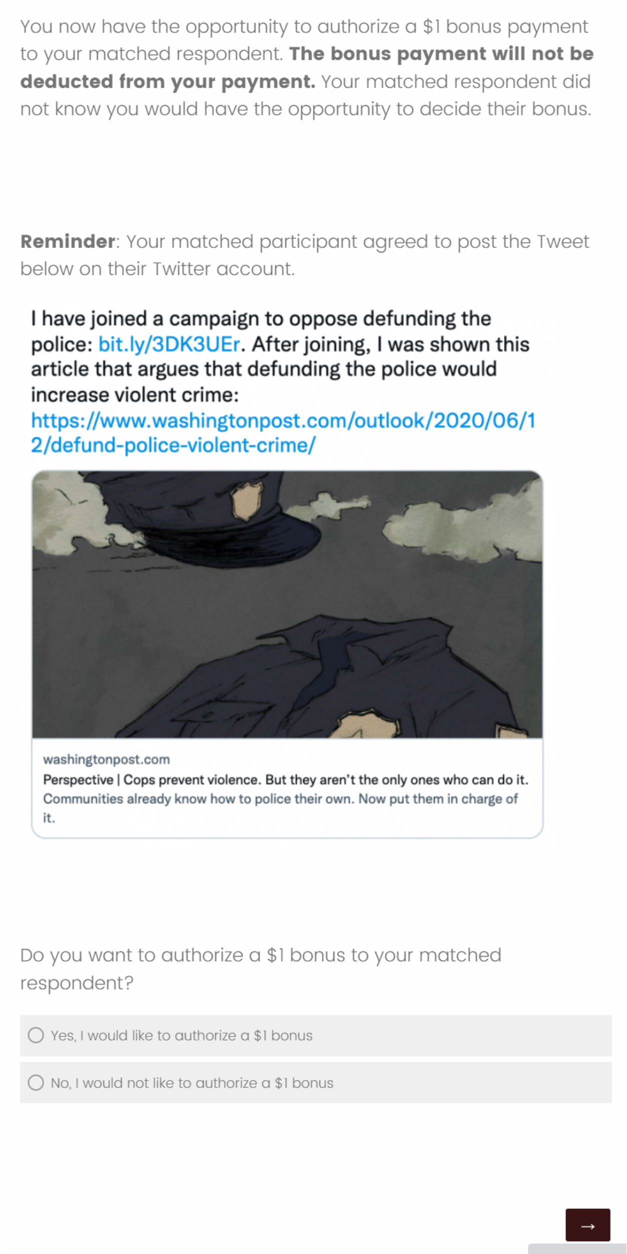
I have joined a campaign to oppose defunding the police: bit.ly/3DK3UEr. After joining, I was shown this article that argues that defunding the police would increase violent crime: https://www.washingtonpost.com/outlook/2020/06/12/defund-police-violent-crime/

washingtonpost.com
Perspective | Cops prevent violence. But they aren't the only ones who can do it.
Communities already know how to police their own. Now put them in charge of it.

Do you think that your matched participant chose to donate $5 to the National Association for the Advancement of Colored People (NAACP)?

○ Yes, I think my matched respondent chose to donate

○ No, I think my matched respondent **did not** choose to donate



You now have the opportunity to authorize a $1 bonus payment to your matched respondent. **The bonus payment will not be deducted from your payment.** Your matched respondent did not know you would have the opportunity to decide their bonus.

Reminder: Your matched participant agreed to post the Tweet below on their Twitter account.

I have joined a campaign to oppose defunding the police: bit.ly/3DK3UEr. After joining, I was shown this article that argues that defunding the police would increase violent crime: https://www.washingtonpost.com/outlook/2020/06/12/defund-police-violent-crime/

washingtonpost.com

Perspective | Cops prevent violence. But they aren't the only ones who can do it.

Communities already know how to police their own. Now put them in charge of it.

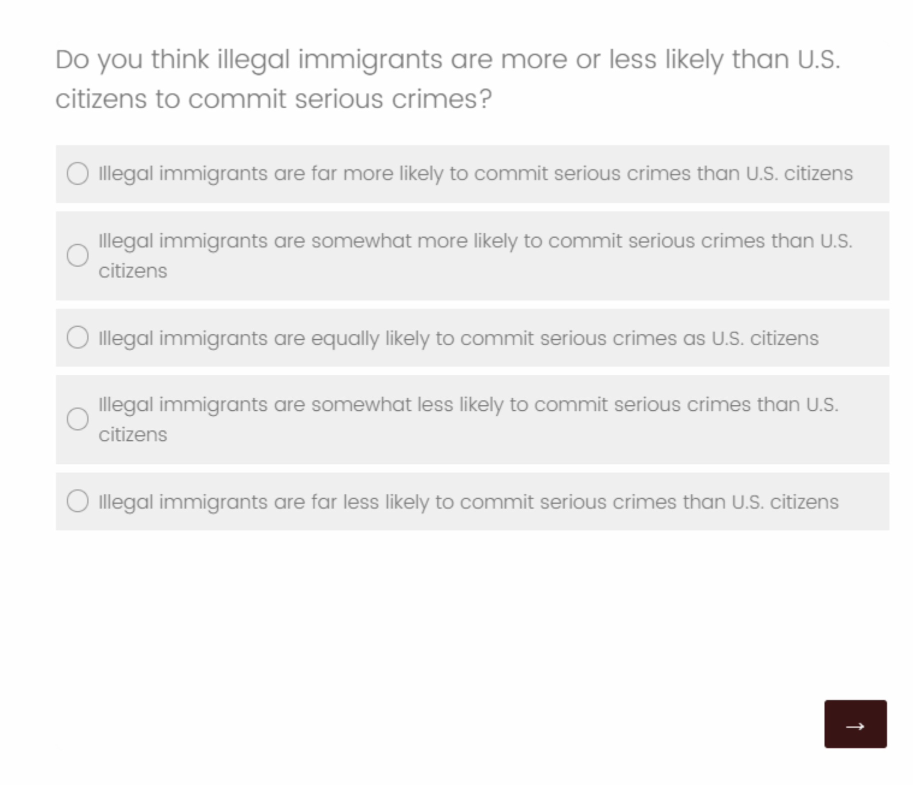
Do you want to authorize a $1 bonus to your matched respondent?

○ Yes, I would like to authorize a $1 bonus

○ No, I would not like to authorize a $1 bonus

E.16 Auxiliary Experiment 8: Persuasion experiment – Republicans

E.16.1 Pre-treatment beliefs



Do you think illegal immigrants are more or less likely than U.S. citizens to commit serious crimes?

○ Illegal immigrants are far more likely to commit serious crimes than U.S. citizens

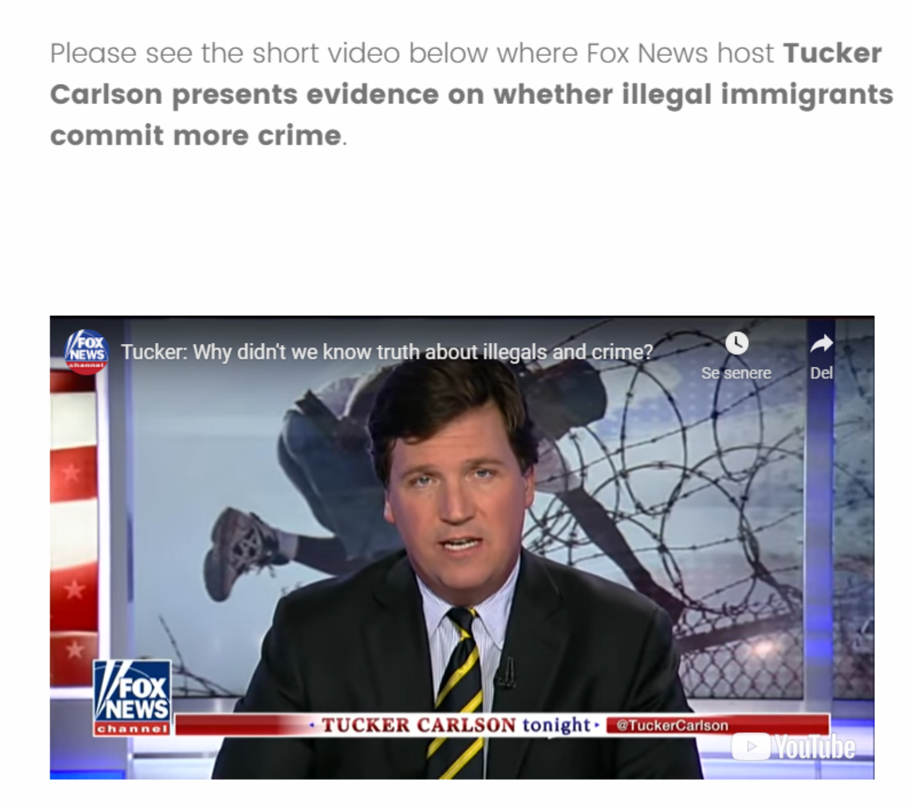
○ Illegal immigrants are somewhat more likely to commit serious crimes than U.S. citizens

○ Illegal immigrants are equally likely to commit serious crimes as U.S. citizens

○ Illegal immigrants are somewhat less likely to commit serious crimes than U.S. citizens

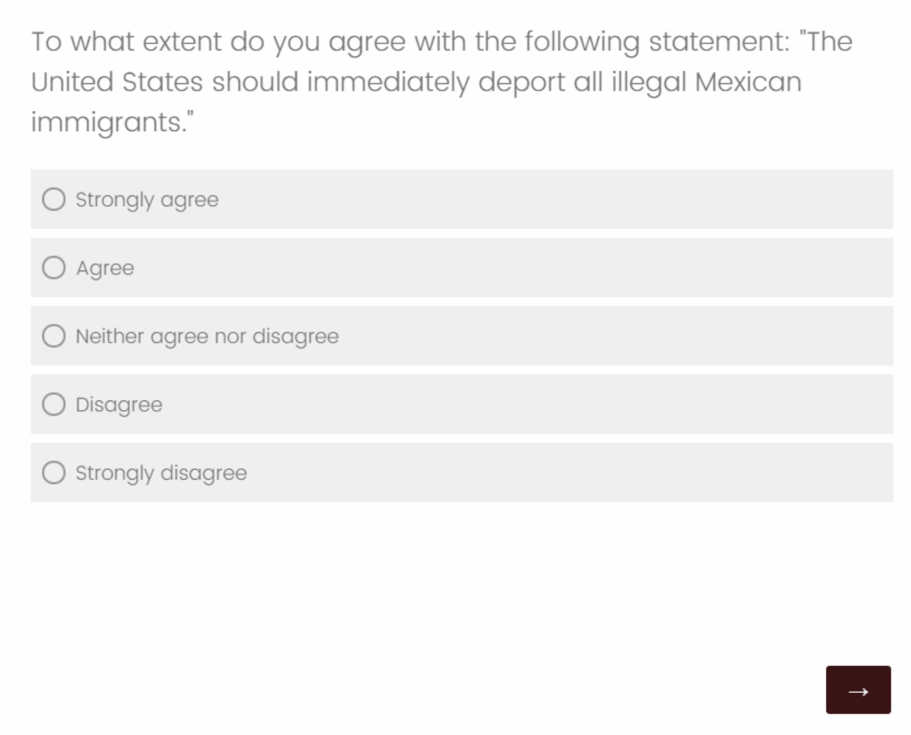
○ Illegal immigrants are far less likely to commit serious crimes than U.S. citizens

E.16.2 Information treatment (only shown to respondents in the treatment group)



Please see the short video below where Fox News host **Tucker Carlson presents evidence on whether illegal immigrants commit more crime**.

E.16.3 Post-treatment outcomes



To what extent do you agree with the following statement: "The United States should immediately deport all illegal Mexican immigrants."

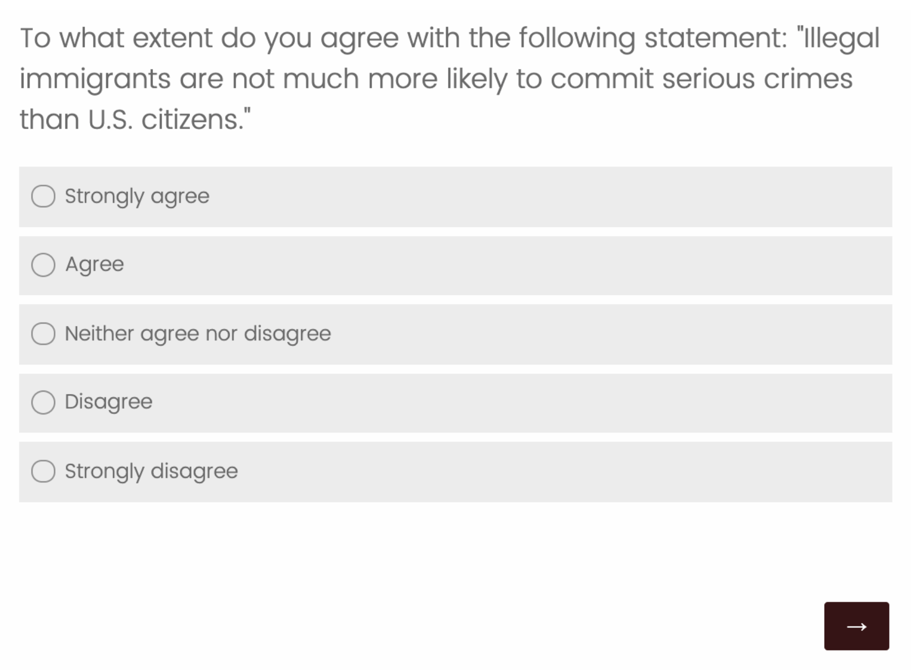
○ Strongly agree

○ Agree

○ Neither agree nor disagree

○ Disagree

○ Strongly disagree



To what extent do you agree with the following statement: "Illegal immigrants are not much more likely to commit serious crimes than U.S. citizens."

○ Strongly agree

○ Agree

○ Neither agree nor disagree

○ Disagree

○ Strongly disagree