Justifying Dissent*

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

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

*This paper supersedes an earlier draft circulated under the title "Disguising Prejudice: Popular Rationales as Excuses for Intolerant Expression." We thank Ben Enke, Davide Cantoni, Ed Glaeser, Daniel Gottlieb, Emir Kamenica, Ro'ee Levy, Ross Mattheis, Nathan Nunn, Pietro Ortoleva, Peter Schwardmann, Marco Tabellini, Joel Van der Weele, David Yang, Noam Yuchtman, Florian Zimmermann, and numerous seminar participants for very helpful suggestions. We thank Danil Fedchenko, Takuma Habu, Hrishikesh Iyengar, Melisa Kurtis, and Alison Zhao for 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 and the Humanities and Social Sciences Research Ethics Committee at the University of Warwick.

[†]University of Chicago and NBER, bursztyn@uchicago.edu
[‡]Kellogg School of Management and NBER, g-egorov@kellogg.northwestern.edu
[§]University of Bergen and CESifo, ingar.haaland@uib.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 either by a majority or by a small but vocal minority. 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 present experiments exploring the power and potential limitations of rationales in facilitating the expression of dissent. Across the political spectrum, dissent is often expressed — and suppressed — on social media, where rationales from both mainstream and fringe sources proliferate and where people often face large social costs from expressing controversial opinions. Motivated by a simple theoretical framework, we experimentally examine the expression and interpretation of dissent on social media in two contentious and policy-relevant domains: liberals' opposition to defunding the police and conservatives' support for deporting illegal immigrants.

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

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

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 to 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: respondents are 12 percentage points more likely to post the Tweet in the *Cover* condition than in the *No Cover* condition. In a placebo experiment with an identical design, but with a Tweet expressing support for a non-stigmatized cause, we find no difference between posting rates in the *Cover* and *No Cover* conditions, suggesting effects are indeed driven by (anticipated) changes in the stigma associated with dissenting expression rather than some other independent effect of the treatment. An additional experiment in which respondents describe the considerations on their mind when posting potentially controversial content corroborates the importance of the cover effect of rationales.

We conduct a second experiment, again with liberal respondents, to examine how 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, 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 31 percent) and are 7 percentage points less likely to deny their matched

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

participant the $1 bonus (relative to a *No Cover* mean of 47 percent). In an additional inference experiment, we show that slightly lowering the credibility of the rationale by removing mention of the academic background of the article's author largely eliminates the treatment effects, highlighting the limits of rationales in shaping inference.

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 audience. 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 the 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 56 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 11 percent) and are 7 percentage points less likely to deny their matched participant the bonus (relative to a *No Cover* mean of 80 percent). Additional experiments with large and representative samples in a somewhat more stylized setting confirm that the availability of rationales increases respondents' willingness to publicly express anti-immigrant sentiments (by donating to an anti-immigration organization) and shapes an audience's inference about underlying motives.

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. By challenging the credibility of rationales or explicitly linking them to stigmatized positions, a vocal group, even a vocal *minority*, can silence a majority — even when the "silent majority" knows they are a majority.

Related literature Our paper builds on a theoretical literature on the effects of social image concerns on economic and moral decision-making. Most closely related to our work is Bénabou et al. (2020), which presents a model of the production and circulation of arguments justifying actions on the basis of morality. We also build on a growing empirical literature studying the effect of social image concerns on political and economic outcomes.[3] Relative to existing work, the key contribution of this paper is to characterize, both theoretically and empirically, how rationales shape the expression of dissent by providing a "social cover," thus lowering the social costs of expression. Our paper is thus related to laboratory evidence on strategic communication used to

[3]These outcomes include moral behavior, as in Ariely et al. 2009; Lacetera and Macis 2010; Ewers and Zimmermann 2015; voting, as in DellaVigna et al. 2017; tax evasion, as in Perez-Truglia and Troiano 2018; Besley et al. 2019; identity choice, as in Jia and Persson 2019; campaign donations, as in Perez-Truglia and Cruces 2017; educational investments, as in Bursztyn and Jensen 2015; and labor market choices, as in Bursztyn et al. 2017.

justify public charitable donations (Foerster and van der Weele, 2021).[4]

Our work also relates to a small literature on political correctness (Morris, 2001; Golman, 2020) and to work examining how social norms govern public behavior more generally (Bénabou and Tirole, 2006, 2011b; Ali and Lin, 2013; Kuran, 1997). Braghieri (2022) shows that publicly expressed views, which may be affected by political correctness norms, are less informative than private views. Like some of this previous work (Bursztyn et al., 2020a,b), our paper examines how previously-stigmatized public behavior can become socially acceptable, but it differs both conceptually and in its implications for equilibrium expression. Conceptually, we isolate the social cover effect of rationales, which shapes the audience's beliefs about a dissenter's motives, from the persuasive effects of the rationale either on the dissenter or the audience. We show that the social cover provided by rationales increases the public expression of dissent by lowering its social cost, as rationales make public actions less informative about dissenters' underlying type. Practically, the mechanism allows even views that are *privately unpopular* — such as conspiracy theories or extreme statements about certain minorities — to be publicly expressed in equilibrium. At the same time, it can enable a minority to silence majority positions, even in the absence of misperceptions about people's views.

More generally, our work connects to a vast literature on the effects of media and propaganda on political and cultural behavior, such as anti-minority violence (e.g. Yanagizawa-Drott 2014; Enikolopov and Petrova 2015; Adena et al. 2015). This literature, examining persuasion in field settings, often finds substantial effects (e.g. Caprettini et al. 2022) — in contrast to the relatively small effects of persuasion documented in survey experiments in which stated views are private (see, for example, Haaland et al. 2021). While there are a number of plausible explanations for this discrepancy, our paper proposes one possible mechanism: widespread propaganda creates *common knowledge* of rationales and thus a "social amplifier" that magnifies rationales' effect on public

[4]Several laboratory studies show that "moral wiggle room" can have substantial effects on behavior: see, for example, Dana et al. (2007); Golman et al. (2017); Saccardo and Serra-Garcia (2020); Golman et al. (2016); Lazear et al. (2012); Hamman et al. (2010); Exley (2016); Cunningham and de Quidt (2016). Because decisions in these settings are anonymous, these findings can be understood through a behavioral model of self-signaling, as in Bénabou and Tirole (2011a). Our work holds this channel constant by exposing respondents to the same private information set, and we instead examine signaling vis-a-vis others, developing a revealed-preference approach that allows us to study image concerns associated with a natural form of expression — posting on social media — before a natural audience — people's actual social media followers. Theoretically, our argument follows the tradition of signal jamming (Holmstrom, 1982; Fudenberg and Tirole, 1986), where individual's action (the use of rationale in our case) makes the individual's characteristics more difficult to infer. Our work is also related to work on "excuses" in psychology, such as Langer et al. (1978), which finds that subjects are more likely to comply with a request (to jump a line to make Xerox copies) justified by a reason. Our paper probes the mechanisms by which some rationales shift social inference about underlying motives — and others fail to do so — and thus sheds light on how rationales can enable dissent.

behavior by generating social cover. 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): authoritarian populists are often highly skilled at generating social cover for exclusionary policies targeting minority groups.

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.

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.

2.1 Setup

The society A 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. $h(\cdot)$, and that it satisfies the monotone hazard rate property ($\frac{h(x)}{1-H(x)}$ is increasing in x, which is satisfied, e.g., for the normal and uniform distributions). To avoid corner cases, we assume that t_i has full support on the real line.

A citizen $i \in A$ is given a chance to publicly state support for change (decision $d_i = 1$) before

an audience. Doing so results in expressive benefit B but social cost S, so $U_i(d_i = 1) = B - S$. We assume that

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

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. For simplicity, we assume that stigma is linear in the audience's posterior about citizen i's type:

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

Lastly, the utility from inaction ($d_i = 0$) is normalized to 0: $U_i(d_i = 0) = 0$.[5]

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 $B = \beta(w_0 + t_i)$. Citizen i makes the decision holding his social cost S fixed. Therefore, he 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 satisfying the condition

$$\tau = \frac{\gamma}{\beta}\mathbb{E}\left(t_i \mid t_i > \tau\right) - w_0. \tag{1}$$

Generally speaking, the threshold need not be unique due to strategic complementarity: if not only extreme right but also moderate types choose action $d_i = 1$, 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 $t_i > \tau$ choose $d_i = 1$ and those with $t_i < \tau$ choose $d_i = 0$.*

[5] 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.

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.

2.3 Persuasive Rationales

Suppose that citizen i, prior to making the decision, received an informative signal $s = w + \varepsilon$, where $\varepsilon \sim \mathcal{N}\left(0, \sigma_\varepsilon^2\right)$. His posterior expectation about w 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 $s > w_0$. Now, 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. Now that more moderate people choose $d_i = 1$, the social cost of doing so is lower: intuitively, publicly supporting C is no longer a sign of extremism. Of course, a decrease in S will prompt even more people to choose $d_i = 1$ (a "social amplifier"). In the end, we have the following characterization of the new equilibrium.

Proposition 2. *Suppose that citizen i makes his decision after receiving informative signal $s > w_0$. This citizen then has a higher posterior about w than the prior, and the ex ante probability that citizen i chooses $d_i = 1$ is higher. The equilibrium social cost S is lower with signal s than without.*

2.4 Polarizing Rationales

In reality, individuals are often presented with the same evidence, but the evidence is interpreted differently. This may be due to differences in background knowledge, cognitive limitations, or behavioral biases, among other factors. For example, different individuals may pick up different arguments from a long article. Alternatively, some credulous individuals may take the text at face value, while others know the bias of a particular journalist or the news outlet and update accordingly.

To study this possibility, we assume 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 $w_l < w_0$). We prove the following result.

Proposition 3. *Suppose that*

$$\mu\left(H\left(\tau\right)-H\left(\tau-\left(w_h-w_0\right)\right)\right)\geq\left(1-\mu\right)\left(H\left(\tau+\left(w_0-w_l\right)\right)-H\left(\tau\right)\right), \tag{2}$$

where τ 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 $d_i = 1$ will go down in equilibrium, and more people will 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 S now contains some types with $t_i < \tau$ (moderates who got a high signal s_h) and lacks some types with $t_i > \tau$ (extremists who got a low signal s_l). As long as 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 face a lower social cost for doing so. This result is not knife-edge, and would apply even if somewhat more people are dissuaded.

Taken together, Propositions 2 and 3 imply that while informative and persuasive evidence can reduce the social cost of a stigmatized public action and lead to more people doing it, evidence that dissuades as many people as it persuades can also be effective at that, due to the social inference problem that such evidence creates. Put differently, for a rationale to be effective it does not have to be persuasive, as long as it hinders inference about the motives for a public action.

3 Opposition to Defunding the Police

The primary experiments in this paper examine the expression of dissent on social media, which we view as an ideal setting for several reasons. First, 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).

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 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 that (1) holds the *persuasion* and *anticipated persuasion* effects constant while varying only the availability of a social cover; (2) measures respondents' revealed-preference willingness to express dissent on their Twitter account; (3) avoids respondents actually posting these Tweets; and (4) avoids explicit deception. We discuss the ethical considerations underlying all experimental designs in Appendix D.

3.1.2 Sample and experimental design

Sample composition We conducted our pre-registered Experiment 1 in October 2021 with a sample of 1,122 Democrats and Independents.[6] As explained below, this resulted in a final sample for analysis of 529 respondents. Given the need for respondents to (1) have an active Twitter account and (2) be willing to log into the survey using their Twitter account, as described below, recruiting respondents to participate in this experiment was more difficult than we anticipated. To reach our pre-registered minimum of 500 complete responses, we recruited respondents from both Luc.id and CloudResearch, two survey providers widely used in the social sciences (Wood and Porter, 2019; Litman et al., 2017).[7] Our final sample is well-balanced on observables across treatment arms (Appendix Table B2).

Twitter login 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 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, 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.

[6] Our experiment was pre-registered in the AEA RCT registry under ID AEARCTR-0008432. The full set of experimental instructions is included in Appendix E.1.

[7] Our final analysis sample consists of 382 respondents from Lucid and 147 respondents from CloudResearch. The two estimates using the samples individually are very similar in size (12.6 p.p. on CloudResearch vs 11.3 p.p. on Luc.id) and statistically indistinguishable.

Information provision and campaign support 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.[8] 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, leaving us with 529 remaining respondents. These respondents are presented with the article again and informed that they can spend as long as they wish reading it.

Twitter post 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.

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 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

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

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 be exposed to the article, since it is linked in the Tweet) across conditions. By employing a one-word manipulation, we also hold other potential confounds, such as the length of the Tweet, fixed across conditions.

3.1.3 Results

Figure 2 displays the results, which we also show in regression table form in Table 1. 57% of respondents authorize the Tweet in the *No Cover* condition compared to 70% of respondents in the *Cover* condition ($p < 0.01$). These effects are stable to the inclusion of demographic and partisan 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).[9] This relatively large effect underscores the importance of the cover effect in driving the expression of dissent.

3.1.4 Ruling out alternative explanations

Placebo One potential concern is that respondents are more willing to schedule the *Cover* Tweet ("Before I joined the campaign...") than the *No Cover* Tweet ("After I joined the campaign...") for reasons unrelated to the availability of the social cover. For example, respondents might think the "before" wording in the *Cover* Tweet sounds more natural than the "after" wording in the *No Cover* Tweet.

To address this concern, we run a placebo experiment (Auxiliary Experiment 2)[10] with the same design and manipulation, but in a different, non-stigmatized context — conservation of the Amazon rainforest — and with a different rationale — an article reporting a new study which finds that over 10,000 species are at risk due to deforestation in the Amazon. Panel A of Figure 3 and Appendix Table B5 show no significant difference between posting rates in the *Cover* and *No Cover* conditions. The difference in effect sizes between the defunding experiment and the placebo experiment is large in magnitude and significant at the 1% level, suggesting effects are indeed driven by (anticipated) changes in the stigma associated with dissenting expression rather than some other

[9] 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.2 for details and Appendix E.5 for experimental instructions.

[10] See Appendix B.3 for details and Appendix E.6 for experimental instructions.

independent effect of the before/after wording.

The placebo results also deliver additional insight into the effect sizes documented in the main experiment. The difference in the fraction of respondents authorizing the post in the *No Cover* treatment, *conditional on privately joining the campaign* — 83% in the placebo experiment, compared to 57% in the main experiment — constitutes suggestive evidence for the existence of (perceived) social sanctions for opposing police defunding and suggests that credible rationales may significantly reduce the extent to which these sanctions prevent the public expression of dissent.

Anticipated persuasion While implausible, it remains possible 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. To directly address this concern, we run an auxiliary experiment (Auxiliary Experiment 3) in which 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 their Tweet.[11] Panel B of Figure 3 and Appendix Table B6 show a small and insignificant 1.9 percentage point difference, suggesting that differences in posting rates are not driven by differences in the anticipated persuasiveness of the Tweets.

3.1.5 Direct evidence on social cover mechanism

Finally, we 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.

Experimental design We conducted Auxiliary Experiment 4 with a sample of 400 Democrats with Twitter accounts recruited from Prolific. 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.[12] 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:

[11] See Appendix B.4 for details and Appendix E.7 for experimental instructions.
[12] See Appendix E.8 for experimental instructions.

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 main object of interest is the difference in respondents' explanations between conditions.

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.[13]

(i) "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. 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."

(ii) "Anticipated persuasion" responses mention that the article might persuade others. For instance, 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 other to side and agree with you. Tweet B does this, Tweet A doesn't."

(iii) "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. For example, one

[13] Our categories themselves are mutually exclusive, but a response might fall under multiple categories if the respondent mentions multiple reasons for their choice. We hand-coded the responses in a team of two people blindly to the treatment status.

respondent writes: "I would want others to see this article and know that I have some evidence to back my tweet."

As evidenced by the "Information" example above, 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. 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, $p = 0.311$). More importantly, the high fraction choosing the Tweet with the rationale (whether the *Cover* or the *No Cover* version) over the *Control* Tweet suggests a widespread preference for citing evidence when engaging in dissenting expression.

We next turn to the open-ended text. As shown in Panel C of Figure 3, 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 (54 percent) fall into the "Information" category. However, given the considerations above, these figures likely substantially underestimate the number of respondents who meant to convey concerns relating to social cover.

Focusing on treatment effects across conditions, the one-word manipulation indeed induces substantially more respondents to mention social cover (a 9 percentage point difference, or a 55 percent effect relative to the *No Cover* mean). Consistent with our previous experiment measuring anticipated persuasion, the manipulation appears to have no effect on the probability that respondents mention that their followers will find the article persuasive. Finally, the *Cover* treatment decreases the share of responses in the "Information" category by 6 percentage points (though this difference is not statistically significant). This presumably reflects the fact that some respondents would have

written an "Information" response if assigned to the *No Cover* treatment (e.g. "The article provides context for my decision") but instead wrote a "Social cover" response when assigned to the *Cover* treatment (e.g. "The article explains why I made the decision").

The perceived social costs of dissent in this setting are evidenced by the substantial number of Tweets mentioning some form of social sanction — for example, "The second tweet contains more information and explanation regarding my choice. Rather than being thrown off by misunderstanding the notion that I want to not defund the police, a direct explanation as to why would make people feel less angry, more understanding, and more interested." Importantly, *zero* responses mention concerns about the "before" or "after" wording being misleading, and only two respondents (both in the *Cover* treatment) write that the Tweet with the rationale sounds strange or unnatural. Alternative interpretations of the treatment effects — for example, respondents feeling that perceived pressure by the experimenter (rather than the article itself) will serve as a social cover, or respondents differentially desiring to signal their support of the article's message — are not supported by the open-ended responses. The small fraction of respondents who choose the *Control* Tweet without a rationale generally cite its shorter length as the reason for doing so (for example, "Because it's short and makes the point. I don't think a lot of people would be interested in clicking and reading the article."). Given that the one-word manipulation in Experiment 1 holds the length of the Tweet fixed, preferences for shorter Tweets will not affect our results.

Together, the placebo experiment, the anticipated persuasion experiment, and this experiment eliciting participant's reasoning establish that the treatment effects documented in Experiment 1 are indeed driven by differences in the availability of a social cover.

3.2 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. We now examine empirically how the availability of rationales affects inference about the dissenter's motives and the social sanctions levied upon the dissenter.

3.2.1 Sample and experimental design

Sample composition We conducted our pre-registered Experiment 2 in November 2021 with a sample of Democrats and Independents recruited from Prolific.[14] Our final sample of 1,040

[14]Our experiment was pre-registered in the AEA RCT registry under ID AEARCTR-0005462. The full set of experimental instructions is included in Appendix E.2.

Democrats and Independents is mostly balanced on observables across treatment arms (Appendix Table B7).

Information about matched respondent Figure 4 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...").

Inference and sanctions We begin by asking respondents to respond to the following open-ended question: "Why do you think your matched participant chose to donate to the campaign?" 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.

3.2.2 Results

We estimate statistically and economically significant treatment effects on all three measures of type inference. Panel A of Figure 5 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 2). 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, Panel B of Figure 5 displays the fraction of participants who deny their matched participant a bonus (results reported in regression table form in Panel B, Columns 1–3 of Table 2). 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 2, these estimates are stable to the inclusion of demographic and partisan controls. As implied by our

framework, even respondents who *privately agree* with their matched participant's opposition to defunding the police may choose to levy social sanctions if they believe that the only people who would be comfortable *publicly* expressing such an opinion are prejudiced.

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 p:

$$\chi^2_p = \frac{\left(n^R_p n^{NR}_{\sim p} - n^{NR}_p n^R_{\sim p}\right)^2}{(n^R_p + n^{NR}_p)(n^R_p + n^R_{\sim p})(n^{NR}_p + n^{NR}_{\sim p})(n^R_{\sim p} + n^{NR}_{\sim p})},$$

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. This statistic will be higher when the use of p is more asymmetric across treatment conditions and lower for phrases that are used rarely across both conditions. Appendix Figure B1 plots the top 100 phrases by χ^2 (where the χ^2 statistics corresponding to phrases more characteristic of the *No Cover* condition have been multiplied by -1 to facilitate visualization). 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, "read an article," "convincing," "increase violent crime," "study" — while respondents in the *No Cover* instead use phrases such as "Republican," "racist," and "probably white".[15]

3.2.3 Exploring the role of credibility

How credible must rationales be in order to be effective? Insufficiently credible rationales may fail to shift social inference and thus social sanctions: a dissenter's audience may not believe the dissenter was persuaded by the rationale, and thus the rationale may not provide the dissenter social cover. 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

[15] 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 donate and this pressure led them to do so. No respondents mention this explanation, and inspection of the phrases most characteristic of each condition suggest that this explanation does not drive the treatment effects.

and thus may be willing to dissent in a wider variety of contexts.

To investigate the role of credibility, we run a slightly revised version of Experiment 2 (Auxiliary Experiment 5) with a sample of 506 Democrats and Independents.[16] We adjust the Tweet (in both the *Cover* and *No Cover* condition) to remove the reference to Sharkey's academic credentials and to the scientific evidence underlying the article's claims. The revised 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]

Importantly, the article — published in the reputable *Washington Post* — remains constant, as does every other aspect of the experiment). Qualitatively speaking, then, this change does not represent a dramatic reduction in credibility. Nonetheless, as shown in Panels C and D of Figure 5 and Columns 4–6 of Table 2, the point estimate of the effect of the rationale on both structured measure of inference remains positive, but is much 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.[17] While we are underpowered to conclude that these treatment effects of around 2 percentage points are statistically significantly smaller than the treatment effects of around 7 percentage points estimated using the more credible rationale, the evidence is qualitatively consistent with this slightly less credible rationale being substantially less effective.

The fact that even an article from a highly credible (liberal-leaning) source might fail to facilitate liberals' disagreement with the "politically correct" position illustrates how public dissent can be silenced by a vocal *minority*: 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). To the extent that this phenomenon generalizes, then, it suggests that for politically charged issues, only highly credible rationales may be effective in facilitating liberal dissent. This may thus stifle dissent on issues for which a strong scientific consensus does not yet exist. 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

[16] See Appendix E.9 for experimental instructions.

[17] As shown in Appendix Table B8, our sample is balanced on observables across treatment arms.

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. 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.

4 Support for Deporting Illegal Immigrants

Our next set of 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 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.[18]

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

4.1 Experiment 3: Rationales and Pro-Deportation Expression

4.1.1 Sample and experimental design

Sample composition We conducted our pre-registered Experiment 3 in March 2021 with a sample of Republicans and Independents.[19] 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 517 respondents. Our sample is balanced on observables across treatment arms (Appendix Table C1).

Experimental design Our experimental design is broadly similar to that of Experiment 1; we provide a diagram in Figure 6. 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 embedded clip from *Tucker Carlson Tonight* 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 then provided with the link 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.

Respondents who join the campaign are informed that one component of the campaign involves circulating a petition on Twitter calling for illegal Mexican immigrants to be deported. We show them a screenshot of the Tweet and ask them if they are willing to schedule it to be posted on their account. As in Experiment 1, we inform respondents that all Tweets will be posted all at once if and when we have surveyed people in all US counties, that this is a common tactic used to make campaigns trend on Twitter, and that we will delete all identifying information by no later than August 1, 2021. Again as in Experiment 1, because we target fewer respondents than the number of US counties, we ensure that Tweets will never be posted. We discuss the ethical considerations

[19] Our experiment was pre-registered in the AEA RCT registry under ID AEARCTR-0007379. The full set of experimental instructions is included in Appendix E.3.

underlying our design in Appendix D.

Respondents in the *Cover* condition are asked whether they would like to schedule 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]

The key experimental manipulation is similar to that of Experiment 1: 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.[20] 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).[21]

4.1.2 Results

Figure 7 displays the results, which we also show in regression table form in Table 3. We again find an economically and statistically significant cover effect: 48% of respondents in the *No Cover* condition authorize the Tweet, while 65% of respondents in the *Cover* condition authorize the Tweet ($p < 0.01$, a 0.35 standard deviation effect). This estimate is stable to the inclusion of

[20] 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.

[21] 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.

demographic and partisan controls. The fact that the 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 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.[22]

4.2 Experiment 4: Interpretation of Pro-Deportation Rationale

We next 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

Sample composition We conducted our pre-registered Experiment 4 in November 2021 with a sample of 1,082 Democrats and Independents recruited from Prolific.[23] 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 C3).

Experiment 4 Experiment 4 follows the structure of Experiment 2; Figure 4 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..."). Respondents then respond to the following open-ended question: "Why do you think your matched participant chose to donate to the campaign?". Subsequently, they guess whether their matched participant authorized a $5 donation to the US Border Crisis Children's Relief Fund (an organization that

[22] In our pre-registered Auxiliary Experiment 6 designed to measure the persuasiveness of the rationale, we find mixed evidence for persuasive effects on private opinions; see Appendix C.2 for details and Appendix E.10 for experimental instructions.

[23] Our experiment was pre-registered in the AEA RCT registry under ID AEARCTR-0005462. The full set of experimental instructions is included in Appendix E.4.

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.[24]

4.2.2 Results

Figure 8 reveals statistically and economically significant treatment effects. In Panel A of Figure 8, we display the fraction of participants in the *Cover* and *No Cover* condition who believe their matched participant donated to the pro-immigrant organization (results are also reported in regression table form in Panel A of Table 4). 8.5% of respondents in the *No Cover* condition believe their matched participant donated, compared to 13.4% of respondents in the *Cover* condition ($p < 0.01$). Similarly, Panel B of Figure 8 displays the fraction of participants who deny their matched participant a bonus (results reported in regression table form in Panel B of Table 4). 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 Table 4, these estimates are stable to the inclusion of demographic and partisan controls. We plot results from our analysis of open-ended text in Appendix Figure C1 using the same procedure described in Section 3.2.2. As in Experiment 2, respondents in the *Cover* condition are substantially more likely to use words referencing the rationale — "watched a video," "right wing media," "link" — whereas respondents in the *No Cover* condition mention phrases such as "Republican," "extremist," and "biased".

4.3 Generalizing Findings

The social media setting of Experiments 1 and 3 affords a highly natural setting and audience and a real-stakes outcome — and is of itself a context of policy relevance — but there are two potential concerns about external validity. First, Twitter users still comprise a relatively small and selected fraction of the population, particularly among Republicans (Wojcik and Hughes, 2019). Second, our requirement that respondents grant our "Tweetability" app permissions to schedule posts on their Twitter account likely induces selection into our experiment. While this selection does not affect the internal validity of Experiments 1 and 3, it might affect the extent to which the results generalize to the broader population. To address these concerns, we present an additional experiment (Auxiliary Experiment 7) that sacrifices some of the naturalness of Experiments 1 and 3 for a large and representative sample, while retaining a revealed-preference measure of respondents'

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

willingness to publicly express dissent. We briefly describe the design and results here, relegating details to Appendix C.4 and a discussion of ethical considerations to Appendix D.

We conduct this experiment in two waves in January 2020 and September 2020 with a total of 5,736 Republican and Independent respondents. We outline the design in Appendix Figure C2. All respondents are first told about the preliminary findings of an unpublished study (Lott, 2018) claiming that immigrants commit more crime than US citizens. Respondents are informed that they will have the opportunity to authorize a $1 donation to Fund The Wall, an organization seeking to construct the proposed US–Mexico border wall, and that we will post their individual donation decision on our website. To vary the availability of a social cover, we tell respondents assigned to the *No Cover* treatment that the web page will state that "all participants were surveyed before Dr. Lott's study was published in an academic journal." In the *Cover* treatment, we instead tell respondents the web page will state that "all participants were shown the preliminary findings from Dr. Lott's study before deciding whether or not to donate to Fund The Wall."

Appendix Table C4 shows that the availability of a social cover increases respondents' willingness to authorize the public donation: 46% of respondents authorize the donation in the *No Cover* condition, while 52% of respondents do so in the *Cover* condition ($p < 0.01$).[25] Furthermore, examining heterogeneity in treatment effects by the 2016 vote share of the respondents' county, we find substantially larger treatment effects in more liberal areas: that is, the availability of a social cover seems to be more important when the audience contains a larger fraction of people who disagree with dissenter's position. This heterogeneity also helps mitigate concerns about experimenter demand effects driving the results.

In a further experiment (N=3,047; Auxiliary Experiment 8), we find using a similar design as in Experiments 2 and 4 that the availability of this rationale affects interpretation of the motives underlying the decision to donate to Fund The Wall: donors with a rationale are seen as less intolerant than those without a rationale.[26] Together, these robustness experiments suggest that the availability of rationales may be an important determinant of the expression and interpretation of dissent in contexts beyond social media.

[25] The smaller effect size in this experiment relative to our Twitter experiments may reflect the fact that we did not screen out respondents who did not privately support the cause.

[26] We provide further details in Appendix C.5.

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 or beliefs about social welfare, but they can also be used to justify dissent, shifting an audience's inference about the dissenter's motivations. We explore these mechanisms among both liberal and conservative respondents, focusing primarily 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 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.

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 arguments (rationales) cannot be voiced because they are seen as legitimizing dangerous or undesirable causes, and so anyone who voices such an argument 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 arguments such as correspondence studies — which are typically effective rationales — 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 arguments 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.

Individuals or institutions seeking to eliminate certain forms of public behavior — for better or for worse — may use multiple levers to silence dissenters. One lever, explored in Section 3.2.3, is to undermine the credibility of rationales directly. Another lever is to manipulate the real or perceived correlation between knowledge of a rationale and underlying type, tying the rationale directly to the

stigmatized motive.[27] 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, or because only they follow a fringe public figure who spreads the rationale — the rationale is completely ineffective, as to use it is to reveal one's motives with certainty. For instance, giving our experimental respondents the opportunity to cite One America News Network (an extreme right-wing outlet) rather than Washington Post (a mainstream newspaper) or Tucker Carlson (the most popular opinion news host) would likely eliminate real and perceived social cover. Tactics to manipulate the real or perceived correlation between motive and rationale include censoring certain figures or otherwise disallowing them a public platform (e.g. disinviting campus speakers), or branding particular media sources or speakers as fringe. 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). Further exploring the conditions under which rationales are most effective, and the unifying features of effective rationales, is an important direction for future research.

Political entrepreneurship and populism Successful politicians often base their campaigns on simple messages and policy platforms that resonate with the general public. Populist politicians are particularly skilled at presenting simple explanations for crises, often blaming scapegoats such as elites or minority groups.[28] Our framework can shed light on why some politicians and some appeals are more effective than others. While the persuasive effects of 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, an effect documented for Nazi propaganda in Weimar Republic by Satyanath et al. (2017) (see also Bursztyn et al. 2019 for evidence of social networks propagating extreme views about ethnic minorities in Russia). The strength of this "social amplifier" channel depends not only on the number of individuals who hold stigmatized views, but the number of individuals who *could not express these views* prior to the rationale becoming widespread.

This distinction can provide one explanation for why the Nazis were able to leverage social networks and associations while other parties, including communists, could not: if antisemitism

[27] 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 — credited as the "father of the atomic bomb" — was stripped of his security clearances when political opponents branded his opposition to the development of the hydrogen bomb to alleged Soviet loyalties (Cassidy, 2005).

[28] See Guriev and Papaioannou (2020) for a review on the political economy of populism. Bursztyn et al. (2022b) applies our framework to explore the scapegoating of minorities during economic crises.

was stigmatized, but relatively common and persistent (Voigtländer and Voth, 2012), then Nazi rhetoric blaming Jews for Germany's problems generated a large social amplifier, thereby furthering Nazi views. Blaming elites, on the other hand, was less stigmatized, and thus generated far smaller amplifiers. Interestingly, Cantoni et al. (2019) show that in the 2017 Bundestag election, Alternative fur Deutschland (AfD) captured more votes in places that voted for Nazis in 1933 (consistent with AfD providing right-wing rationales in general) but not in places where antisemitic pogroms were historically more numerous, as AfD does not use any explicitly antisemitic rhetoric.[29]

Fake and misleading news Our findings are also relevant for the debate about the influence of fake and misleading news on society. Some recent studies suggest that their persuasive effect is limited (Allcott and Gentzkow, 2017; Nyhan, 2018), while others suggest that they can be effective at changing 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., 2022a). Our results point, however, to an alternative mechanism through which misleading news can affect public expression. Specifically, fake news can generate a "social amplifier": rationales that plausibly persuade a small subset of the population can change public behavior among a much larger fraction of the population, increasing their willingness to express otherwise-stigmatized views by leveraging fake news as a rationale. Interestingly, in Barrera et al. (2020), subjects exposed to fake news were not only more willing to support an extreme candidate (Marine Le Pen), but also were unlikely to change their opinion after being exposed to fact-checks — even though these fact-checks improved factual knowledge. This evidence is difficult to explain by the persuasive power of fake news alone, but it is consistent with the role of fake news as rationales: fake and misleading news can generate social cover for individuals to express extreme views, and debunking does not eliminate social cover as long as the fact-check can be plausibly dismissed.

This insight has implications for debunking fake news spread online and offline. 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 users' timelines (e.g., because

[29] Germany has strict laws on anti-Semitic speech which criminalize Holocaust denial or "incitement to hatred" (Volksverhetzung). Our logic shows how that laws can effectively prevent certain rationales from being used, even though they cannot be fully effective against, for example, dog whistles, "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" (Goodin and Saward, 2005; Haney-López, 2014; Grosjean et al., 2020), or disguising taste-based discrimination (Becker, 1957) as statistical discrimination (Phelps, 1972; Arrow, 1973).

a friend shared it). The former initiative decreases the persuasive effect of fake news for a user who seeks to spread it, while the latter decreases the anticipated persuasiveness of the rationale. Yet because these experiments have occurred only among a small fraction of users, people have a ready-made social cover when sharing fake news: they can credibly claim that they were not warned the news was fake.[30]

Our results highlight the potential importance of eliminating social cover: ensuring that the audience *knows that the poster knew the news had been debunked* and nonetheless chose to post it. A simple path would be to scale the debunking experiments to the entire userbase, thus generating common knowledge that all users are warned before posting fake news. Because the general equilibrium results of such a change differ significantly from the partial equilibrium results, current estimates of the effects of debunking on users' propensity to share fake news may substantially understate the true 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 credibility is lacking (for example, as a result of past mistakes) fake and misleading news retains its power to generate social cover for the expression of stigmatized views.

Rationales beyond politics: "Acting White" Austen-Smith and Fryer (2005) model negative stigma associated with educational investment — "Acting White" — which might arise if choosing to study signals that the individual has limited social opportunities, or if choosing to study signals that the individual is weakly attached to their social group. In such settings, providing monetary incentives for exerting educational effort (as discussed in Levitt et al. 2016) might provide students with a rationale, allowing them to attribute educational investments not to academic interest but rather to the incentive. For similar reasons, in these settings, cold-calling after asking a question might be preferable to allowing students to volunteer answers. Understanding the use of rationales to generate social cover in educational and other non-political settings is an important area for future investigation.

[30] Indeed, both Twitter and Facebook's fact-checking efforts have been widely criticized for a lack of transparency, and it is thus certain that most users lack information about how the platforms fight misinformation (Nyhan, 2017).

References

Acemoglu, Daron, Georgy Egorov, and Konstantin Sonin, "A Political Theory of Populism," *The 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," *The Quarterly Journal of Economics*, 2015, *130* (4), 1885–1939.

Alesina, Alberto, Armando Miano, and Stefanie Stantcheva, "Immigration and redistribution," Working Paper 24733, National Bureau of Economic Research, 2019.

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," Technical Report, mimeo, Columbia University 2021.

Ariely, Dan, Anat Bracha, and Stephan Meier, "Doing Good or Doing Well? Image Motivation and Monetary Incentives in Behaving Prosocially," *American Economic Review*, 2009, *99* (1), 544–555.

Arrow, Kenneth, "The Theory of Discrimination," in Albert Rees and Orley Ashenfelter, eds., *Discrimination in Labor Markets*, Princeton, New Jersey: Princeton University Press, 1973.

Austen-Smith, David and Roland G. Fryer, "An Economic Analysis of "Acting White"," *The 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.

Becker, Gary S., *The Economics of Discrimination*, University of Chicago Press, 1957.

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," *The Quarterly Journal of Economics*, 2011, *126* (2), 805–855.

_ and _ , "Laws and Norms," Technical Report w17579, National Bureau of Economic Research 2011.

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

Besley, Timothy, Anders Jensen, and Torsten Persson, "Norms, Enforcement, and Tax Evasion," Technical Report, National Bureau of Economic Research 2019.

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

Bursztyn, Leonardo, Aakaash Rao, Chris Roth, and David Yanagizawa-Drott, "Opinions as Facts," Technical Report 2022.

_ , 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 K Haaland, Aakaash Rao, and Christopher Roth, "Scapegoating During Crises," Technical Report 2022.

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

_ , Thomas Fujiwara, and Amanda Pallais, "'Acting Wife': Marriage Market Incentives and Labor Market Investments," *American Economic Review*, 2017, *107* (11), 3288–3319.

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

Caprettini, Bruno, Marcel Caesmann, Hans-Joachim Voth, and David Yanagizawa-Drott, "Going Viral: Propaganda, Persuasion and Polarization in 1932 Hamburg," Technical Report, Working Paper 2022.

Cassidy, David, *J. Robert Oppenheimer and the American Century*, Pi Press, 2005.

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.

Cunningham, Tom and Jonathan de Quidt, "Implicit Preferences Inferred from Choice," *Available at SSRN 2709914*, 2016.

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.

DellaVigna, Stefano, John A. List, Ulrike Malmendier, and Gautam Rao, "Voting to Tell Others," *The Review of Economic Studies*, 2017, *84* (1), 143–181.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *arXiv:1810.04805 [cs]*, 2019. arXiv: 1810.04805.

Ekins, Emily, "Poll: 62% of Americans Say They Have Political Views They're Afraid to Share," *Cato Institute*, 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 of *Handbook of Media Economics*, North-Holland, 2015, pp. 687–700.

Ewers, Mara and Florian Zimmermann, "Image and Misreporting," *Journal of the European Economic Association*, 2015, *13* (2), 363–380.

Exley, Christine L., "Excusing Selfishness in Charitable Giving: The Role of Risk," *The 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.

Fudenberg, Drew and Jean Tirole, "A "signal-jamming" theory of predation," *The RAND Journal of Economics*, 1986, pp. 366–376.

Gennaioli, Nicola and Andrei Shleifer, "What Comes to Mind*," *The 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.

_ , Bryan Kelly, and Matt Taddy, "Text as Data," *Journal of Economic Literature*, 2019, *57* (3), 535–74.

Golman, Russell, "Acceptable Discourse: Social Norms of Beliefs and Opinions," *Working paper*, 2020.

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

_ , George Loewenstein, Karl Ove Moene, and Luca Zarri, "The Preference for Belief Consonance," *Journal of Economic Perspectives*, 2016, *30* (3), 165–88.

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

Grigorieff, Alexis, Christopher Roth, and Diego Ubfal, "Does Information Change Attitudes Toward Immigrants?," *Demography*, 2020, *57* (3), 1–27.

Grosjean, Pauline A., Federico Masera, and Hasin Yousaf, "Whistle the Racist Dogs: Political Campaigns and Police Stops," SSRN Scholarly Paper ID 3662027, Social Science Research Network, Rochester, NY 2020.

Guriev, Sergei and Elias Papaioannou, "The political economy of populism," *Journal of Economic Literature*, 2020.

Haaland, Ingar and Christopher Roth, "Labor market concerns and support for immigration," *Journal of Public Economics*, 2020, *191*, 104256.

_ , _ , 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 ; New York: Oxford University Press, 2014.

Holmstrom, Bengt, "Managerial Incentive Problems: A Dynamic Perspective," *Essays in Economics and Management in Honor of Lars Wahtbeck (Helsinki: Swedish School of Economics)*, 1982.

Hopkins, Daniel J, John Sides, and Jack Citrin, "The Muted Consequences of Correct Information about Immigration," *The Journal of Politics*, 2019, *81* (1), 315–320.

Jia, Ruixue and Torsten Persson, "Individual vs. Social Motives in Identity Choice: Theory and Evidence from China," Technical Report, National Bureau of Economic Research 2019.

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.

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), 635–642.

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," *The 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," SSRN Scholarly Paper ID 3496903, Social Science Research Network, Rochester, NY 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.

Liu, Qi, Matt J. Kusner, and Phil Blunsom, "A Survey on Contextual Embeddings," *arXiv:2003.07278 [cs]*, 2020. arXiv: 2003.07278.

Lott, John R., "Undocumented Immigrants, U.S. Citizens, and Convicted Criminals in Arizona," Working Paper, Social Science Research Network, 2018.

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, "Why the Fact-Checking at Facebook Needs to Be Checked," *The New York Times*, 2017.

_ , "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," SSRN Scholarly Paper ID 3797290, Social Science Research Network, Rochester, NY 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.

Perez-Truglia, Ricardo and Ugo Troiano, "Shaming Tax Delinquents," *Journal of Public Economics*, 2018, *167*, 120–137.

Phelps, Edmund S., "The Statistical Theory of Racism and Sexism," *American Economic Review*, 1972, *62* (4), 659–661.

Saccardo, Silvia and Marta Serra-Garcia, "Cognitive Flexibility or Moral Commitment? Evidence of Anticipated Belief Distortion," *CESifo Working Paper*, 2020.

Sanh, Victor, Lysandre Debut, Julien Chaumond, and Thomas Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," *arXiv:1910.01108 [cs]*, 2020. arXiv: 1910.01108.

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.

Science Panel for the Amazon, "Amazon Assessment Report 2021," *Science Panel for the Amazon*, 2021.

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," *The Quarterly Journal of Economics*, 2012, *127* (3), 1339–1392.

Wojcik, Stefan and Adam Hughes, "Sizing Up Twitter Users," *Pew Research Center's Report*, 2019.

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," *The Quarterly Journal of Economics*, 2014, *129* (4), 1947–1994.

Figures

Figure 1: Experiment 1: flow of dissent design

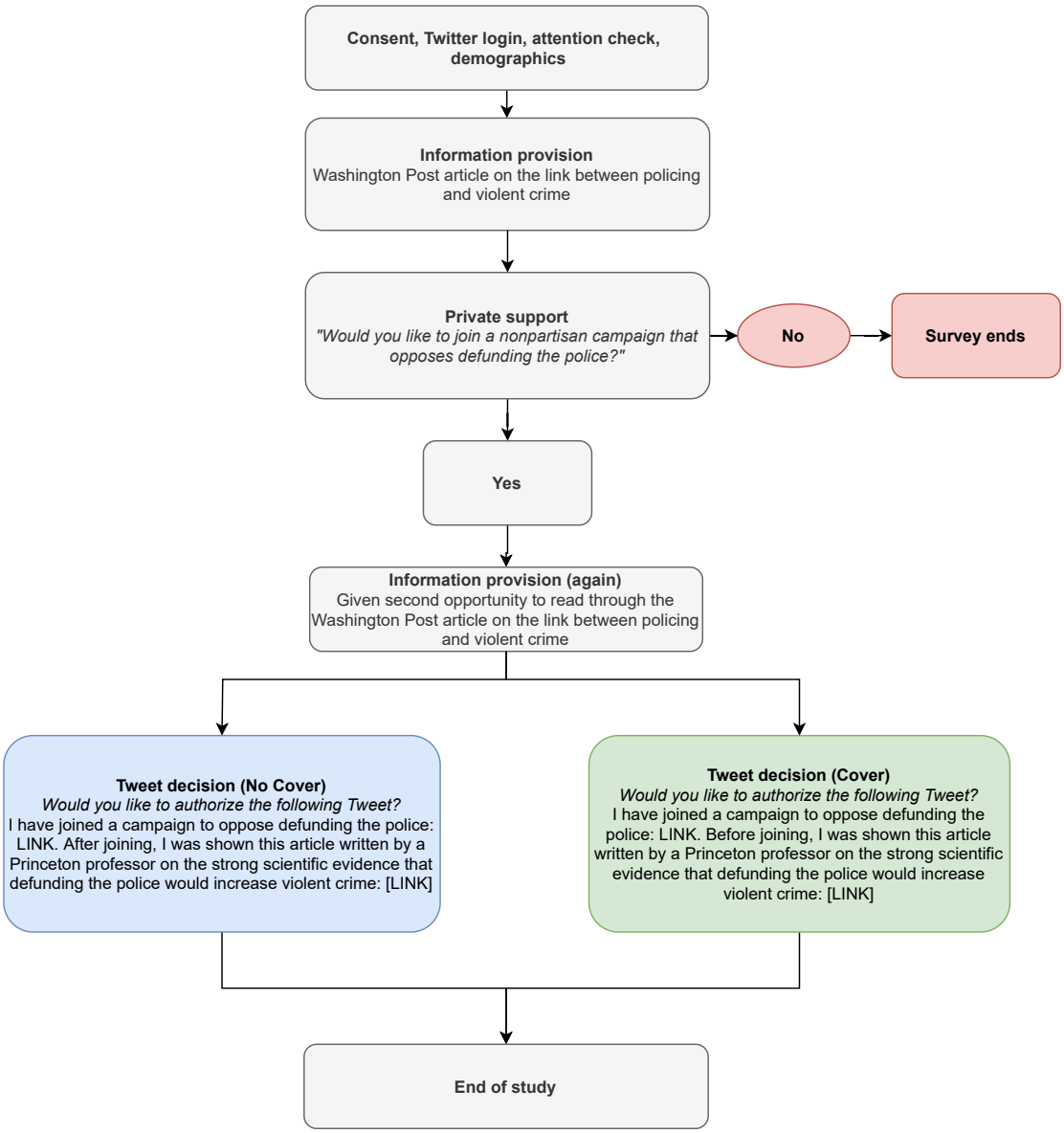
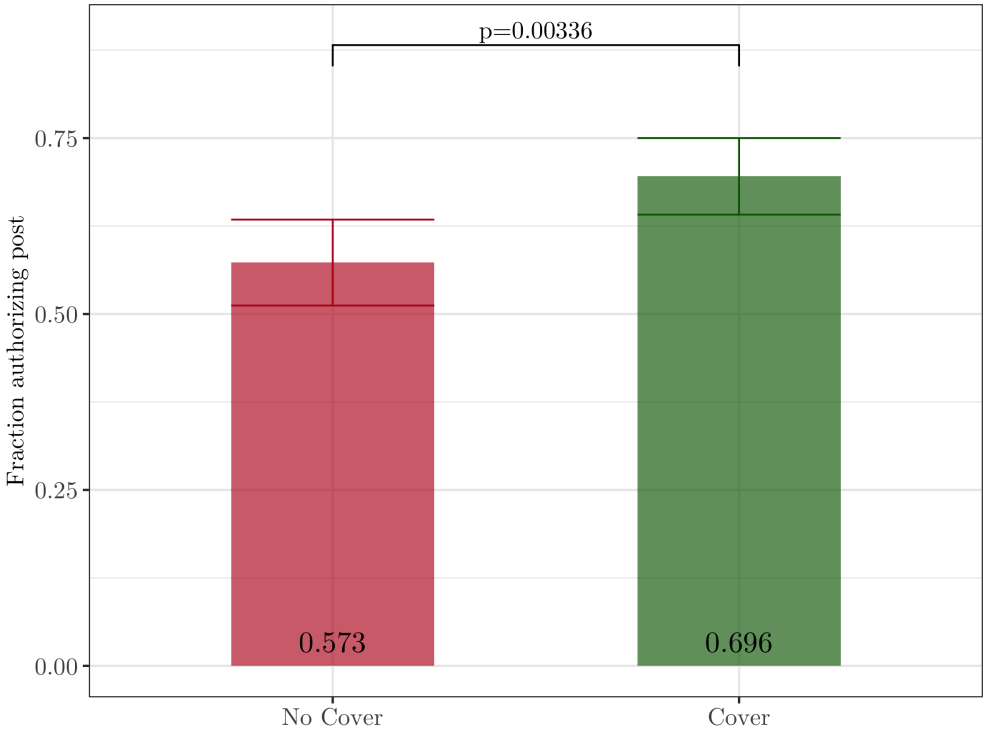
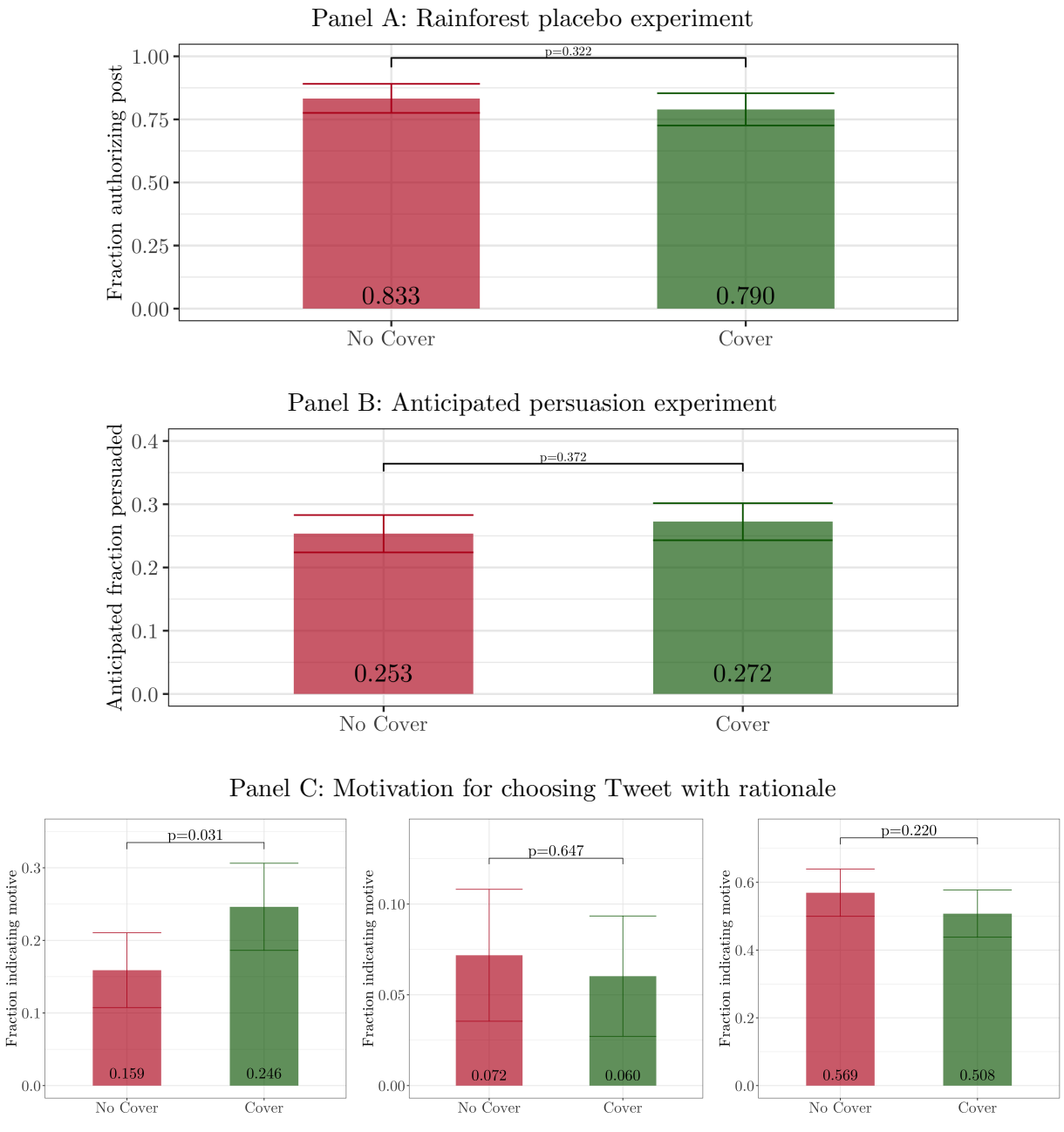


Figure 2: Experiment 1: willingness to post anti-defunding Tweet



Notes: Figure displays the fraction of respondents authorizing the Tweet indicating their opposition to the movement to defund the police, separately by experimental condition. Error bars indicate 95% confidence intervals. p-values obtained from a two-sample t-test of equality of means.

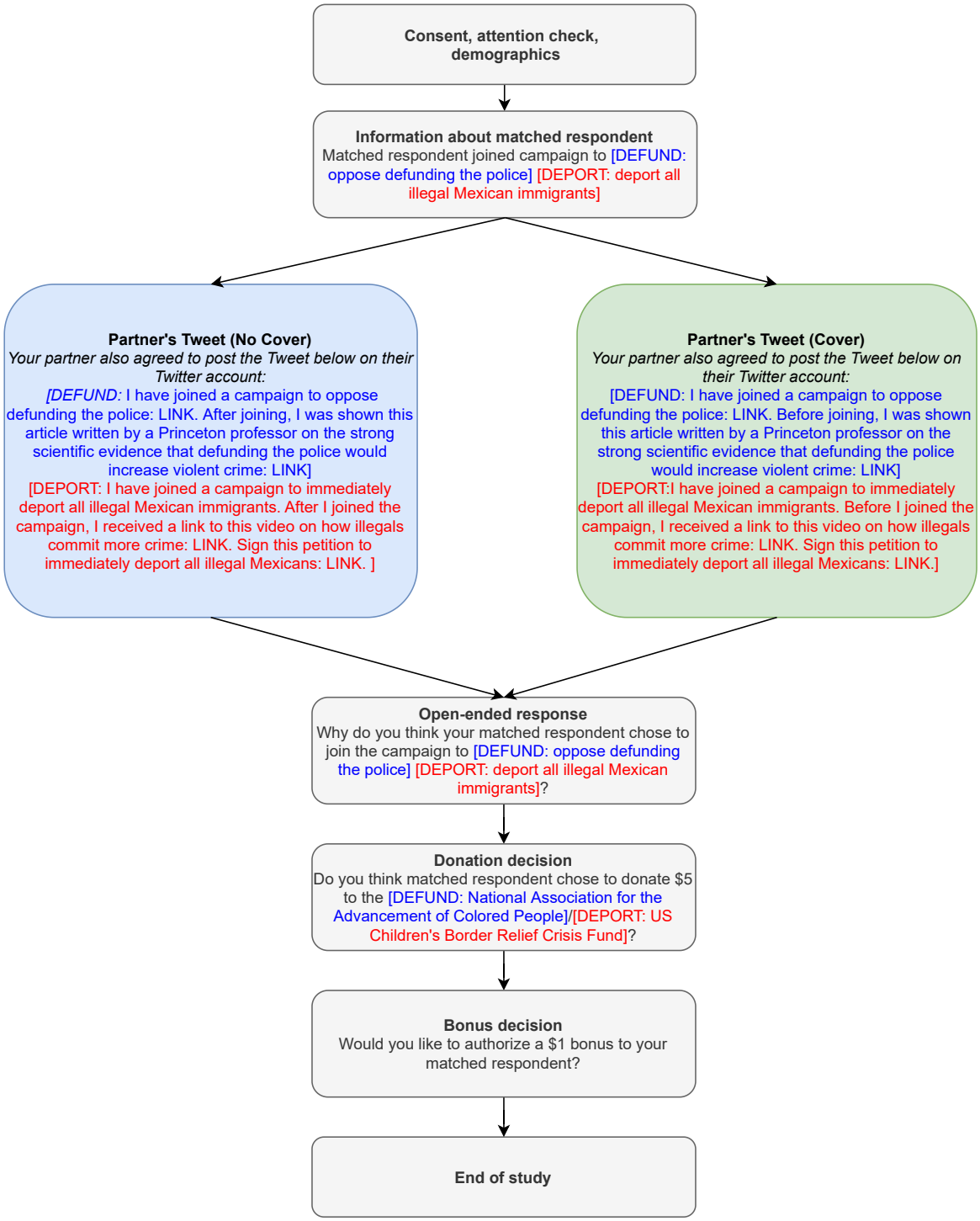
Figure 3: Experiment 1: ruling out alternative interpretations



(a) Social Cover (b) Anticipated persuasion (c) Information

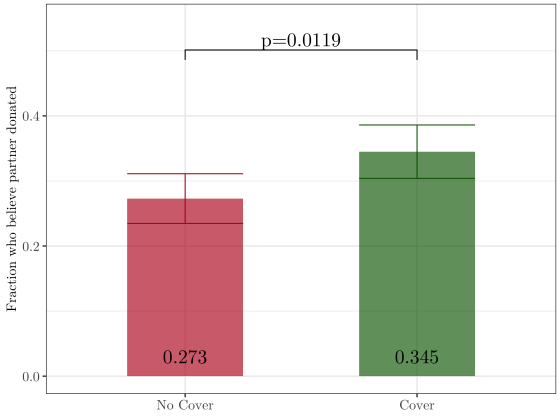
Notes: Panel A of Figure 3 displays the fraction of participants who authorize the Tweet in each condition of the placebo experiment, described in Section 3.1.4. Panel B displays the mean of respondents' guesses as to the fraction of their followers who would be persuaded by the Tweet to join the campaign, elicited in the anticipated persuasion experiment described in Section 3.1.4. Panel C displays the fraction of respondents who mention each of the three motives when choosing which Tweet to post, elicited in the open-ended text experiment described in Section 3.1.5. Error bars indicate 95% confidence intervals. p-values obtained from a two-sample t-test of equality of means.

Figure 4: Experiments 2 and 4: flow of inference design

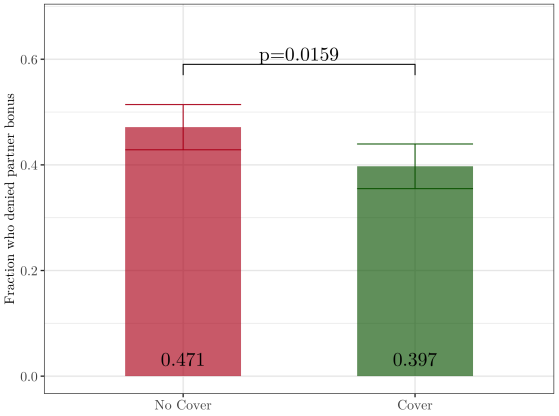


Notes: Experiments 2 and 4 have identical structures, so we present both experiments jointly. Blue text corresponds to Experiment 2, studying opposition to the movement to defund the police; red text corresponds to Experiment 4, studying support for immediately deporting all illegal Mexican immigrants.

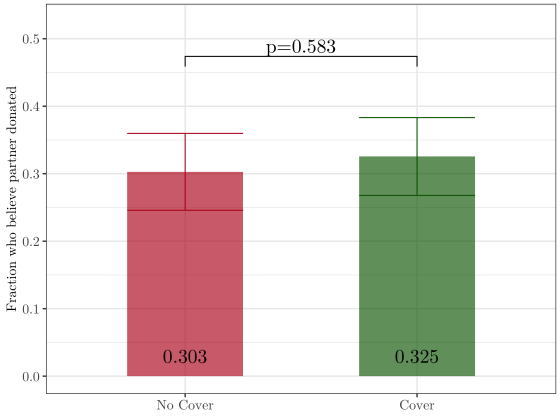
Figure 5: Experiment 2: Interpretation of anti-defunding Tweet



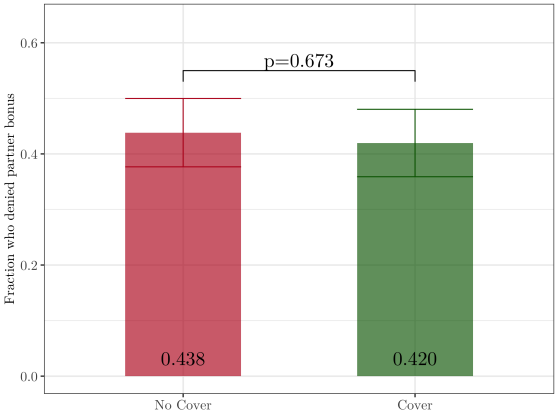
(a) High credibility: beliefs about NAACP donation



(b) High credibility: denial of bonus



(c) Low credibility: beliefs about NAACP donation



(d) Low credibility: denial of bonus

Notes: Panels A and C present the fraction of respondents who believe their matched participant donated to the NAACP (a pro-black organization). Panels B and D present the fraction of respondents who deny their matched participant a $1 bonus. Panels A and B present results from the (high-credibility) main experiment; Panels C and D present results from the lower-credibility experiment. Error bars indicate 95% confidence intervals. p-values obtained from a two-sample t-test of equality of means.

Figure 6: Experiment 3: design

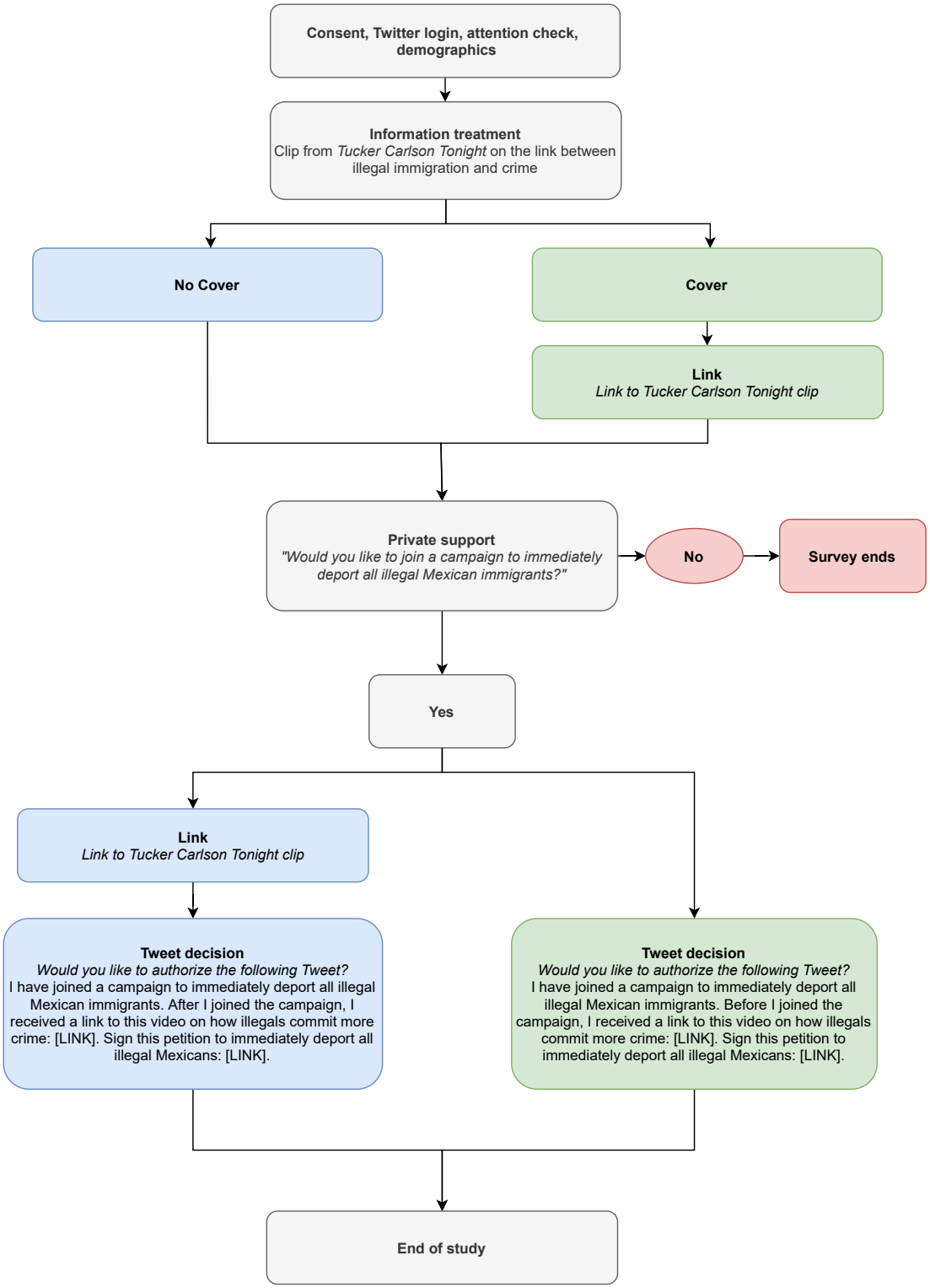
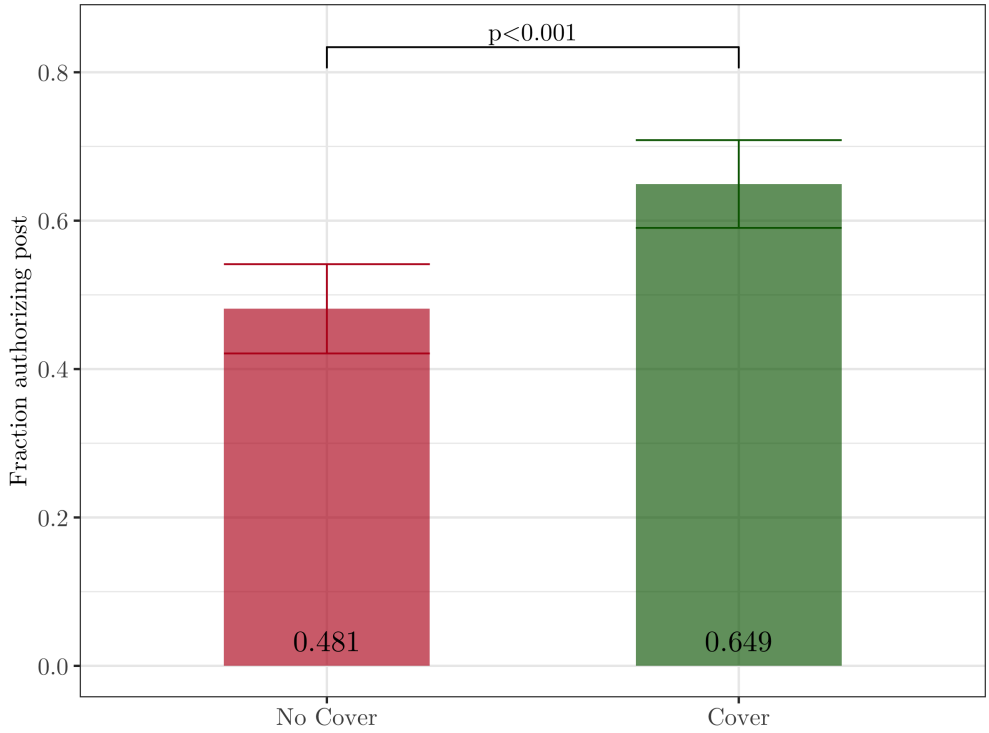
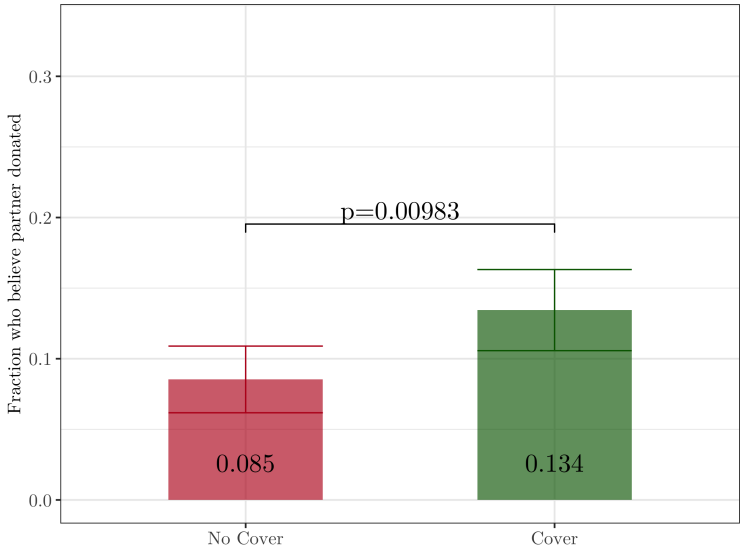


Figure 7: Experiment 3: willingness to post pro-deportation Tweet

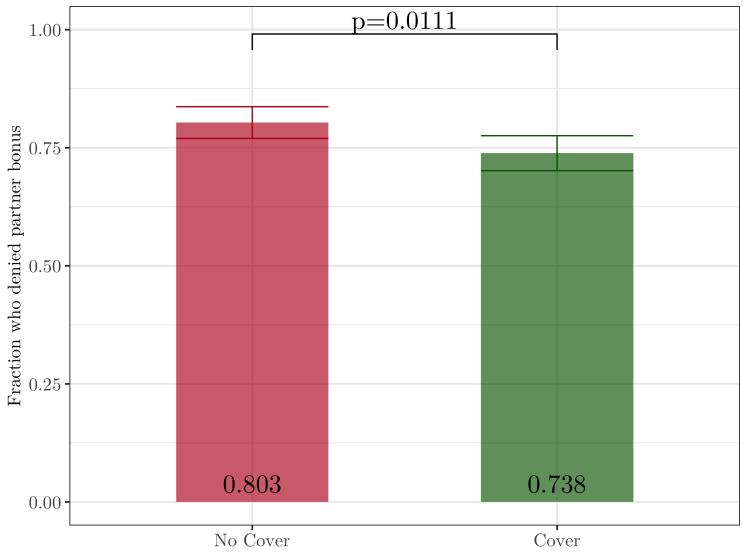


Notes: Figure displays the fraction of respondents authorizing the Tweet indicating their support for immediately deporting all illegal Mexican immigrants. Error bars indicate 95% confidence intervals. p-values obtained from a two-sample t-test of equality of means.

Figure 8: Experiment 4: Interpretation of pro-deportation Tweet



(a) Beliefs about USBCCRF donation



(b) Denial of bonus

Notes: Panel A presents the fraction of respondents who believe their matched participant donated to the US Border Crisis Children's Relief Fund. Panel B presents the fraction of respondents who deny their matched participant 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: Experiment 1: Willingness to post anti-defunding Tweet

	Dependent variable:		
	Anti-defunding Tweet		
	(1)	(2)	(3)
Cover	0.123***	0.118***	0.120***
	(0.042)	(0.042)	(0.042)
DV mean	0.637	0.637	0.637
DV std. dev.	0.481	0.481	0.481
Observations	529	529	529
Demographic controls	No	Yes	Yes
Partisan controls	No	No	Yes

Notes: The dependent variable is an indicator taking value 1 if the respondent chose to schedule the post. Demographic controls include age, age squared, a set of race indicators, a Hispanic indicator, a male indicator, a set of education indicators. Partisan controls include indicators for "Very conservative", "Conservative", "Neither liberal nor conservative" (omitted), "Liberal", and "Very liberal". Robust standard errors are reported.

Table 2: Experiment 2: Inference about and social sanctions toward matched anti-defunding respondent

	High credibility			Low credibility		
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A:	*Belief partner donated*					
Cover	0.072**	0.072**	0.067**	0.023	0.023	0.019
	(0.029)	(0.029)	(0.029)	(0.041)	(0.042)	(0.042)
DV mean	0.309	0.309	0.309	0.314	0.314	0.314
DV std. dev.	0.462	0.462	0.462	0.465	0.465	0.465
Panel B:	*Denied bonus to partner*					
Cover	−0.074**	−0.074**	−0.067**	−0.019	−0.028	−0.015
	(0.031)	(0.031)	(0.030)	(0.044)	(0.044)	(0.043)
DV mean	0.435	0.435	0.435	0.429	0.429	0.429
DV std. dev.	0.496	0.496	0.496	0.495	0.495	0.495
Observations	1,040	1,037	1,036	506	506	506
Demographic controls	No	Yes	Yes	No	Yes	Yes
Partisan controls	No	No	Yes	No	No	Yes

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 1–3 report results for the main (high-credibility) experiment; Columns 4–6 report results for the lower-credibility experiment. Demographic controls include age, age squared, a set of race indicators, a Hispanic indicator, a male indicator, a set of education indicators. Partisan controls include indicators for "Very conservative", "Conservative", "Neither liberal nor conservative" (omitted), "Liberal", and "Very liberal". Robust standard errors are reported.

Table 3: Experiment 3: Willingness to post pro-deportation Tweet

	Dependent variable:		
	Anti-immigrant Tweet		
	(1)	(2)	(3)
Cover	0.168***	0.175***	0.173***
	(0.043)	(0.043)	(0.043)
DV mean	0.563	0.563	0.563
DV std. dev.	0.497	0.497	0.497
Observations	517	517	517
Demographic controls	No	Yes	Yes
Partisan controls	No	No	Yes

Notes: The dependent variable is an indicator taking value 1 if the respondent chose to schedule the post. Demographic controls include age, age squared, a set of race indicators, a Hispanic indicator, a male indicator, a set of education indicators. Partisan controls include indicators for "Very conservative", "Conservative", "Neither liberal nor conservative" (omitted), "Liberal", and "Very liberal". Robust standard errors are reported.

Table 4: Experiment 4: Inference about and social sanctions toward matched pro-deportation respondent

	(1)	(2)	(3)
Panel A:	*Belief partner donated*		
Cover	0.049***	0.051***	0.048**
	(0.019)	(0.019)	(0.019)
DV mean	0.110	0.110	0.110
DV std. dev.	0.313	0.313	0.313
Panel B:	*Denied bonus to partner*		
Cover	−0.065**	−0.065**	−0.061**
	(0.026)	(0.026)	(0.026)
DV mean	0.771	0.771	0.771
DV std. dev.	0.421	0.421	0.421
Observations	1,082	1,081	1,081
Demographic controls	No	Yes	Yes
Partisan controls	No	No	Yes

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. Demographic controls include age, age squared, a set of race indicators, a Hispanic indicator, a male indicator, a set of education indicators. Partisan controls include indicators for "Very conservative", "Conservative", "Neither liberal nor conservative" (omitted), "Liberal", and "Very liberal". Robust standard errors are reported.

Online Appendix:
Not for publication

Our supplementary material is structured as follows. Appendix A provides proofs of all theoretical results in Section 2. Appendix B provides supporting material for the experiments presented in Section 3. Appendix C provides supporting material for the experiments presented in Section 4. Appendix D 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

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

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

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

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

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

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

The integral may be rewritten as

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

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

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

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

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

$$\begin{aligned} \frac{d}{d\tau}\left(\frac{1 - H(x + \tau)}{1 - H(\tau)} \right) &= \frac{h(\tau)(1 - H(x + \tau)) - h(x + \tau)(1 - H(\tau))}{(1 - H(\tau))^2} \\ &= \frac{1 - H(x + \tau)}{1 - H(\tau)} \left(\frac{h(\tau)}{1 - H(\tau)} - \frac{h(x + \tau)}{1 - H(x + \tau)} \right) \leq 0, \end{aligned} \tag{3}$$

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

$$G(\tau) = -w_0, \tag{4}$$

where

$$G(\tau) = \tau - \frac{\gamma}{\beta}\mathbb{E}(t_i \mid t_i > \tau). \tag{5}$$

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 the equation (4) has a unique solution. 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 $w_1 > w_0$. From the proof of Proposition 1, the new equilibrium again takes the form of a threshold $\tilde{\tau}$ that satisfies

$$G(\tilde{\tau}) = -w_1,$$

where $G(\cdot)$ is defined in (5). Since $\frac{d}{d\tau}G(\tau) > 0$ and $-w_1 < -w_0$, we have $\tilde{\tau} < \tau$ (and furthermore, since $\frac{d}{d\tau}G(\tau) < 1$, the difference $\tau - \tilde{\tau} > w_1 - w_0$, so the decrease in threshold τ is larger than the increase in w). Now, $\tilde{\tau} < \tau$ implies that the share of citizens choosing $d_i = 1$ has increased: $1 - H(\tilde{\tau}) > 1 - H(\tau)$. Lastly, the social cost is now equal $\gamma\mathbb{E}(t_i \mid t_i > \tilde{\tau}) < \gamma\mathbb{E}(t_i \mid t_i > \tau)$, so it is lower than without the signal s. This completes the proof. ■

A.3 Proof of Proposition 3

We start by establishing the uniqueness of equilibrium in this case.[31] 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 $\bar{\tau}$.

In what follows, we use the following probabilities. We denote

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

[31] 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 make 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, understands the whole signal structure, but does not know which signal citizen i got, and faces the signal decomposition problem as a result.

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 $\bar{S}\left(\bar{\tau}\right)$. 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

$$\bar{\tau}-\frac{1}{\beta}\bar{S}\left(\bar{\tau}\right)=-w_0. \tag{6}$$

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

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

Furthermore,

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

Lastly, we have

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

Now,

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

where we denoted $u=z+w_0-w_h$ and used the calculation (3) from the proof of Proposition 1.

This immediately implies that $\frac{d}{dz}q\left(z+w_0-w_h,z+w_0-w_l\right)\geq 0$. Summing up, we have

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

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 $\frac{d}{dz}\frac{1}{\gamma}\bar{S}\left(z\right)\leq 1$. Now, as in the proof of Proposition 1 this implies that the equation (6) has a unique solution $\bar{\tau}$, which proves the uniqueness of equilibrium in this case.

Let us now show that in this solution, $\bar{\tau}<\tau$ and $\bar{S}\left(\bar{\tau}\right)<S\left(\tau\right)$, where $S\left(\tau\right)=\frac{1}{\gamma}\mathbb{E}\left(t_i\mid t_i>\tau\right)$ is the equilibrium social cost in the absence of any signal, in the unique solution τ. To do this, it is sufficient to show that $\bar{S}\left(\tau\right)<S\left(\tau\right)$. Indeed, this would imply that

$$\tau-\frac{1}{\beta}\bar{S}\left(\tau\right)>\tau-\frac{1}{\beta}S\left(\tau\right)=-w_0,$$

and since $\bar{\tau}$ satisfies (6) and the function $x-\frac{1}{\beta}\bar{S}\left(x\right)$ is increasing, we would get $\bar{\tau}<\tau$. Then we would get

$$\bar{S}\left(\bar{\tau}\right)=\beta\left(\bar{\tau}+w_0\right)<\beta\left(\tau+w_0\right)=S\left(\tau\right),$$

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

In the light of condition (2) and by continuity of $H\left(\cdot\right)$, there exists $\hat{w}_h\in\left(0,w_h\right)$ such that

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

Let $\hat{S}$ denote the value

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

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

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

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

where

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

Since $\rho>0$ and $\frac{1}{\gamma}\hat{S}<\mathbb{E}\left(t_i\mid t_i\in\left(\tau+w_0-w_h,\tau+w_0-\hat{w}_h\right)\right)$ as the former is an expectation taken over values to the right of $\tau+w_0-\hat{w}_h$ 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 $\hat{S} < S(\tau)$. Spelling out $q(\tau + w_0 - \hat{w}_h, \tau + w_0 - w_l)$ and expectations in the definition of $\hat{S}$, we have

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

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

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

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

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

which is positive, because the first term is greater than τ and the second is less than that. Therefore, $\hat{S} < S(\tau)$.

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

Table B1: Overview of Data Collections

Experiment	**Provider**	**Dates**
Panel A: Main Experiments		
Experiment 1: Willingness to post anti-defunding Tweet – Democrats authorizing Twitter access (N=1,122)	Luc.id, Cloudresearch	October 2021
Experiment 2: Interpretation of anti-defunding Tweet – Democrats (N=1,040)	Prolific	November 2021
Experiment 3: Willingness to post pro-deportation Tweet – Republicans authorizing Twitter access (N=1,130)	Luc.id	March 2021
Experiment 4: Interpretation of pro-deportation Tweet – Republicans (N=1,082)	Prolific	November 2021
Panel B: Auxiliary Experiments		
Auxiliary Experiment 1: Persuasiveness of anti-defunding article – Democrats (N=1,008)	Prolific	December 2021
Auxiliary Experiment 2: Placebo: willingness to post pro-conservation Tweet – respondents authorizing Twitter access (N=483)	Luc.id, Cloudresearch	December 2021 and January 2022
Auxiliary Experiment 3: Anticipated persuasiveness of anti-defunding Tweet – Democrats (N=501)	Prolific	November 2021
Auxiliary Experiment 4: Motives underlying the choice – Democrats with Twitter account (N=400)	Prolific	January 2022
Auxiliary Experiment 5: Interpretation of lower-credibility anti-defunding Tweet – Democrats (N=506)	Prolific	November 2021
Auxiliary Experiment 6: Persuasiveness of pro-deportation Tweet – Republicans (N=2,012)	Prolific, Lucid	December 2021
Auxiliary Experiment 7 (wave 1): Willingness to donate to anti-immigrant organization – Conservatives (N=4,457)	Luc.id	January 2020
Auxiliary Experiment 7 (wave 2): Willingness to donate to anti-immigrant organization – Conservatives (N=1,299)	Luc.id	September 2020
Auxiliary Experiment 8: Interpretation of anti-immigrant donation – Liberals (N=3,047)	Luc.id	February 2020

Notes: Reported sample sizes for Experiment 1, Experiment 3, and Auxiliary Experiment 2 include respondents who chose not to join the campaigns and therefore are not included in the sample we analyze.

B Anti-Defunding Experiments: Additional Material

B.1 Experiment 1: Additional Figures and Tables

Table B2: Experiment 1: Balance of covariates

	Overall		Cover	No Cover	p-value
	mean	std.dev.	mean	mean	(R=NR)
Age	39.877	15.130	39.293	40.514	0.355
Black	0.214	0.410	0.228	0.198	0.391
Asian	0.076	0.265	0.072	0.079	0.775
White	0.671	0.470	0.670	0.672	0.968
Hispanic	0.183	0.387	0.159	0.209	0.138
Male	0.580	0.494	0.572	0.589	0.702
High school diploma	0.975	0.155	0.975	0.976	0.903
Bachelors degree	0.435	0.496	0.417	0.455	0.381

Notes: p-values based on robust standard errors reported.

B.2 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.[32] 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: "Do you think that funding for the police should be increased, decreased, or stay the same?" and "How do you think increasing funding for the police would affect violent crime?". 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 B3 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, with an effect size of around 0.25 standard deviations for the first outcome and 0.12 standard deviations for the second outcome.

B.3 Auxiliary Experiment 2: Rainforest Placebo

B.3.1 Design and results

We conducted this experiment in December 2021 and January 2022 with a sample of 483 Democrats and Independents recruited from Luc.id and CloudResearch. Respondents logged in to the survey with their Twitter accounts using the same procedure as in Experiment 1. The design is similar to that of Experiment 1, but examines a different (non-stigmatized) context: willingness to post

[32]The pre-registration is available in the AEA RCT registry under ID AEARCTR-0008624.

Table B3: Persuasive effects of anti-defunding article

	Dependent variable:					
	Belief			Policy preference		
	(1)	(2)	(3)	(4)	(5)	(6)
Provided article	-0.236^{***}	-0.245^{***}	-0.245^{***}	0.135^{*}	0.121^{*}	0.122^{*}
	(0.056)	(0.055)	(0.055)	(0.071)	(0.068)	(0.068)
DV mean	-0.080	-0.080	-0.080	-0.568	-0.568	-0.568
DV std. dev.	0.902	0.902	0.902	1.133	1.133	1.133
Observations	1,008	1,007	1,007	1,008	1,007	1,007
Demographic controls	No	Yes	Yes	No	Yes	Yes
Partisan controls	No	No	Yes	No	No	Yes

Notes: 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"). Demographic controls include age, age squared, a set of race indicators, a Hispanic indicator, a male indicator, a set of education indicators. Partisan controls include indicators for "Very conservative", "Conservative", "Neither liberal nor conservative" (omitted), "Liberal", and "Very liberal". Robust standard errors are reported.

a Tweet supporting efforts to conserve the Amazon rainforest.[33] Rather than reading an article about the likely effects of defunding the police, respondents read a Reuters article reporting on a study conducted by the Science Panel for the Amazon which finds that over 10,000 species are at risk from deforestation in the Amazon (Science Panel for the Amazon, 2021). The *Cover* Tweet reads:

> 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: [LINK]. Join the campaign and sign the petition: [LINK].

The *No Cover* Tweet is identical, but replaces "Before I joined the campaign..." with "After I joined the campaign...".

Appendix Table B5 shows no significant difference between posting rates in the *Cover* and *No Cover* conditions, and the difference in effect sizes between the defunding experiment and the placebo experiment is large in magnitude (16 percentage points) and significant at the 5% level, suggesting effects are indeed driven by (anticipated) changes in the stigma associated with dissenting expression rather than some other independent effect of the wording.

B.4 Auxiliary Experiment 3: 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

[33] Table B4 shows that our sample is balanced on observables across treatment arms.

Table B4: Rainforest placebo: Balance of covariates

	Overall		Cover	No Cover	p-value
	mean	std.dev.	mean	mean	(R=NR)
Age	39.110	13.244	38.146	40.043	0.201
Black	0.135	0.342	0.172	0.099	0.056
Asian	0.047	0.212	0.025	0.068	0.074
White	0.768	0.423	0.739	0.796	0.226
Hispanic	0.163	0.370	0.166	0.160	0.902
Male	0.473	0.500	0.529	0.420	0.052
High school diploma	0.984	0.124	0.981	0.988	0.628
Bachelors degree	0.392	0.489	0.376	0.407	0.565

Notes: p-values based on robust standard errors reported.

Table B5: Willingness to post pro-conservation Tweet (placebo)

	Dependent variable:		
	Pro-conservation Tweet		
	(1)	(2)	(3)
Cover	−0.044	−0.049	−0.052
	(0.044)	(0.044)	(0.045)
DV mean	0.812	0.812	0.812
DV std. dev.	0.391	0.391	0.391
Observations	319	319	319
Demographic controls	No	Yes	Yes
Partisan controls	No	No	Yes

Notes: The dependent variable is an indicator taking value 1 if the respondent chose to schedule the post. Demographic controls include age, age squared, a set of race indicators, a Hispanic indicator, a male indicator, a set of education indicators. Partisan controls include indicators for "Very conservative", "Conservative", "Neither liberal nor conservative" (omitted), "Liberal", and "Very liberal". Robust standard errors are reported.

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?"

Table B6 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.

Table B6: Anticipated persuasiveness of Tweet

	Dependent variable:		
	Perceived percentage persuaded		
	(1)	(2)	(3)
Cover	1.897	2.190	2.504
	(2.125)	(2.114)	(2.123)
DV mean	26.309	26.309	26.309
DV std. dev.	23.764	23.764	23.764
Observations	501	501	501
Demographic controls	No	Yes	Yes
Partisan controls	No	No	Yes

Notes: The dependent variable is the respondent's guess as to the percentage of their followers who would join the campaign if they saw the Tweet. Demographic controls include age, age squared, a set of race indicators, a Hispanic indicator, a male indicator, a set of education indicators. Partisan controls include indicators for "Very conservative", "Conservative", "Neither liberal nor conservative" (omitted), "Liberal", and "Very liberal". Robust standard errors are reported.

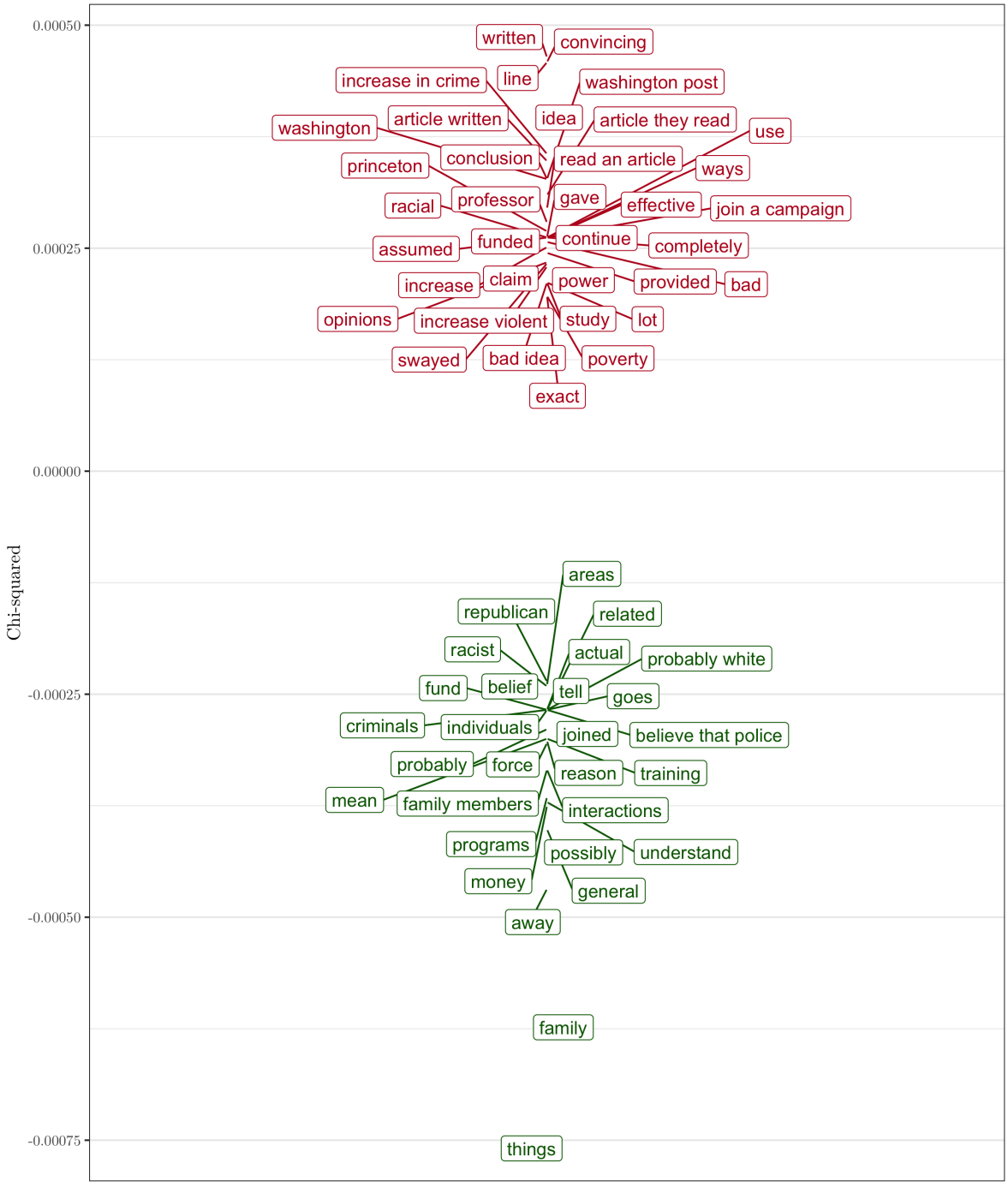
B.5 Experiment 2: Additional Figures and Tables

Table B7: Experiment 2: Balance of covariates

	Overall		Cover	No Cover	p-value
	mean	std.dev.	mean	mean	(R=NR)
Age	30.725	11.258	30.686	30.763	0.912
Black	0.070	0.256	0.086	0.055	0.057
Asian	0.085	0.279	0.089	0.080	0.589
White	0.773	0.419	0.766	0.781	0.563
Hispanic	0.112	0.315	0.093	0.130	0.060
Male	0.374	0.484	0.384	0.365	0.522
High school diploma	0.997	0.054	0.996	0.998	0.552
Bachelors degree	0.572	0.495	0.562	0.582	0.520

Notes: p-values based on robust standard errors reported.

Figure B1: Experiment 2: most distinctive phrases in each condition



Notes: Appendix Figure B1 plots phrases by their associated χ^2 statistic, limiting to the top 100 phrases and multiplying the χ^2 of phrases more characteristic of the "No Cover" condition by -1. The words "article" and "read" have χ^2 values of greater than 0.001 and have been suppressed to facilitate visualization of the remaining phrases.

Table B8: Experiment 2 (lower-credibility): Balance of covariates

	Overall		Cover	No Cover	p-value
	mean	std.dev.	mean	mean	(R=NR)
Age	35.366	14.585	35.275	35.458	0.888
Black	0.053	0.225	0.043	0.064	0.303
Asian	0.132	0.339	0.137	0.127	0.747
White	0.771	0.421	0.773	0.769	0.923
Hispanic	0.107	0.309	0.141	0.072	0.011
Male	0.496	0.500	0.502	0.490	0.789
High school diploma	0.996	0.063	0.992	1.000	0.160
Bachelors degree	0.597	0.491	0.600	0.594	0.884

Notes: p-values based on robust standard errors reported.

C Anti-Immigrant Experiments: Additional Material

C.1 Experiment 3: Additional Tables

Table C1: Experiment 3: Balance of covariates

	Overall		Cover	No Cover	p-value
	mean	std.dev.	mean	mean	(R=NR)
Age	49.422	13.623	48.709	50.094	0.248
Black	0.014	0.116	0.012	0.015	0.762
Asian	0.015	0.124	0.016	0.015	0.934
White	0.952	0.215	0.952	0.951	0.955
Hispanic	0.066	0.248	0.052	0.079	0.214
Male	0.503	0.500	0.490	0.515	0.571
High school diploma	0.994	0.076	0.996	0.992	0.598
Bachelors degree	0.385	0.487	0.343	0.425	0.055

Notes: p-values based on robust standard errors reported.

C.2 Auxiliary Experiment 6: Persuasiveness of Deportation Rationale

We conducted a first pre-registered experiment in December 2021 with a sample of 1,008 Republicans recruited from Prolific.[34] 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" and "The US should immediately deport all illegal Mexican immigrants." We code both questions from -2 ("Strongly disagree") to 2 ("Strongly agree").

Contrary to our pre-registered prediction, we did not estimate a statistically significant effect of viewing the video on either 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. Table C2 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, 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.

[34] The pre-registration is available in the AEA RCT registry under ID AEARCTR-0008624.

Table C2: Persuasive effects of *Tucker Carlson Tonight* video

	Dependent variable:					
	Belief			Policy preference		
	(1)	(2)	(3)	(4)	(5)	(6)
Provided article	-0.133^{*}	-0.123^{*}	-0.123^{*}	0.177^{**}	0.179^{**}	0.180^{**}
	(0.070)	(0.071)	(0.071)	(0.074)	(0.073)	(0.073)
DV mean	0.643	0.643	0.643	0.740	0.740	0.740
DV std. dev.	1.112	1.112	1.112	1.173	1.173	1.173
Observations	1,004	1,002	1,002	1,004	1,002	1,002
Demographic controls	No	Yes	Yes	No	Yes	Yes
Partisan controls	No	No	Yes	No	No	Yes

Notes: 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"). Demographic controls include age, age squared, a set of race indicators, a Hispanic indicator, a male indicator, a set of education indicators. Partisan controls include indicators for "Very conservative", "Conservative", "Neither liberal nor conservative" (omitted), "Liberal", and "Very liberal". Robust standard errors are reported.

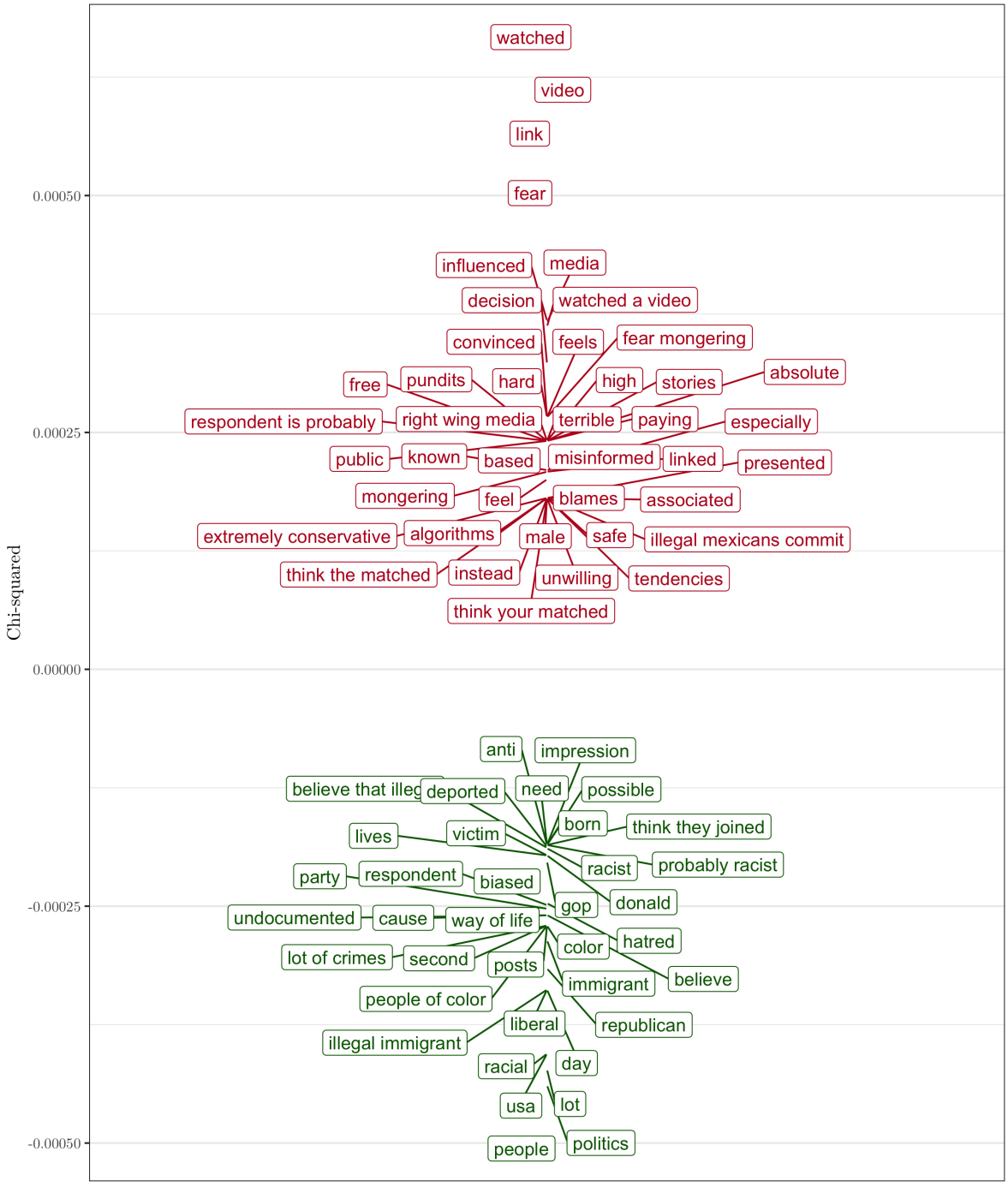
C.3 Experiment 4: Additional Tables

Table C3: Experiment 4: Balance of covariates

	Overall		Cover	No Cover	p-value
	mean	std.dev.	mean	mean	(R=NR)
Age	31.729	12.256	32.408	31.046	0.068
Black	0.069	0.254	0.063	0.076	0.389
Asian	0.100	0.300	0.090	0.109	0.297
White	0.767	0.423	0.785	0.750	0.174
Hispanic	0.118	0.323	0.109	0.128	0.325
Male	0.479	0.500	0.492	0.466	0.392
High school diploma	0.995	0.068	0.994	0.996	0.659
Bachelors degree	0.589	0.492	0.590	0.588	0.939

Notes: p-values based on robust standard errors reported.

Figure C1: Experiment 4: most distinctive phrases in each condition



Notes: Appendix Figure C1 plots phrases by their associated χ^2 statistic, limiting to the top 100 phrases and multiplying the χ^2 of phrases more characteristic of the "No Cover" condition by -1.

C.4 Auxiliary Experiment 7: Anti-Immigrant Expression Among More Representative Sample

The social media setting of Experiments 1 and 3 affords a highly natural setting and real-stakes outcome — and is of itself a context of policy relevance — but there are two potential concerns about external validity. First, Twitter users still comprise a relatively small and selected fraction of the population, particularly among Republicans (Wojcik and Hughes, 2019). Second, our requirement that respondents grant our "Tweetability" app permissions to schedule posts on their Twitter account likely induces selection into our experiment. While this selection does not affect the internal validity of Experiments 1 and 3, it might affect the extent to which the results generalize to the broader population. A third, and related, limitation of these experiments is that we are unable to examine heterogeneity in the effect of the rationale based on the composition of a respondent's audience, both because we cannot observe their audience and because we are insufficiently powered to do so. To address these concerns, this section presents an additional experiment that sacrifices some of the naturalness of Experiments 1 and 3 for a large and representative sample (whose geographic location and thus whose local environment we can observe), while retaining a revealed-preference measure of respondents' willingness to publicly express dissent. We discuss ethical considerations in Appendix D.

C.4.1 Sample and experimental design

Sample composition We conducted Auxiliary Experiment 7 in January 2020 (wave 1) and September 2020 (wave 2) with a sample of Republicans and Independents recruited through Luc.id.[35] Our sample of respondents is broadly representative of Independents and Republicans in the United States (Appendix Table C7) and is well-balanced on observables across treatment arms (Appendix Table C6). The two waves had a largely similar design. The most important differences are that wave 1 included a pure control condition and had a more heavy-handed set of instructions. In the description of the design below, we focus of the leaner set of instructions from wave 2.

Information: Lott study After completing a series of demographic and other background questions, respondents are assigned to two main conditions: *No Cover* and *Cover*.

Specifically, all respondents are first told about the preliminary findings of an unpublished study (Lott, 2018) claiming that immigrants commit more crime than US citizens. Respondents are informed that they will have the opportunity to authorize a $1 donation to Fund The Wall, an organization seeking to construct the proposed US–Mexico border wall, and that we will post their individual donation decision on our website. To vary the availability of a social cover, we tell respondents assigned to the *No Cover* treatment that the web page will state that "all participants were surveyed before Dr. Lott's study was published in an academic journal." In the *Cover* treatment, respondents are instead told that the web page will state that "all participants were shown the the preliminary findings from Dr. Lott's study before deciding whether or not to donate to Fund The Wall." We tell respondents in the *Cover* and *No Cover* conditions about a recent study (Lott, 2018) which argues that undocumented immigrants commit more crimes and more serious crimes than US citizens.[36] This study has been widely covered by the media, including

[35] Wave 1 of the experiment was pre-registered in the AEA RCT registry under ID AEARCTR-0005308.

[36] Wave 1 of study also had a pure *Control* condition in which respondents do not learn about the study. We include a discussion of the results from the *Control* condition below.

The Washington Times, *National Review*, and *Fox News*, and has been repeatedly cited by Trump administration officials. We also truthfully tell our respondents that a number of sources (including a researcher affiliated with the Cato Institute, a libertarian think tank) have recently challenged some of the study's methods, claiming that errors in analysis invalidate its results.

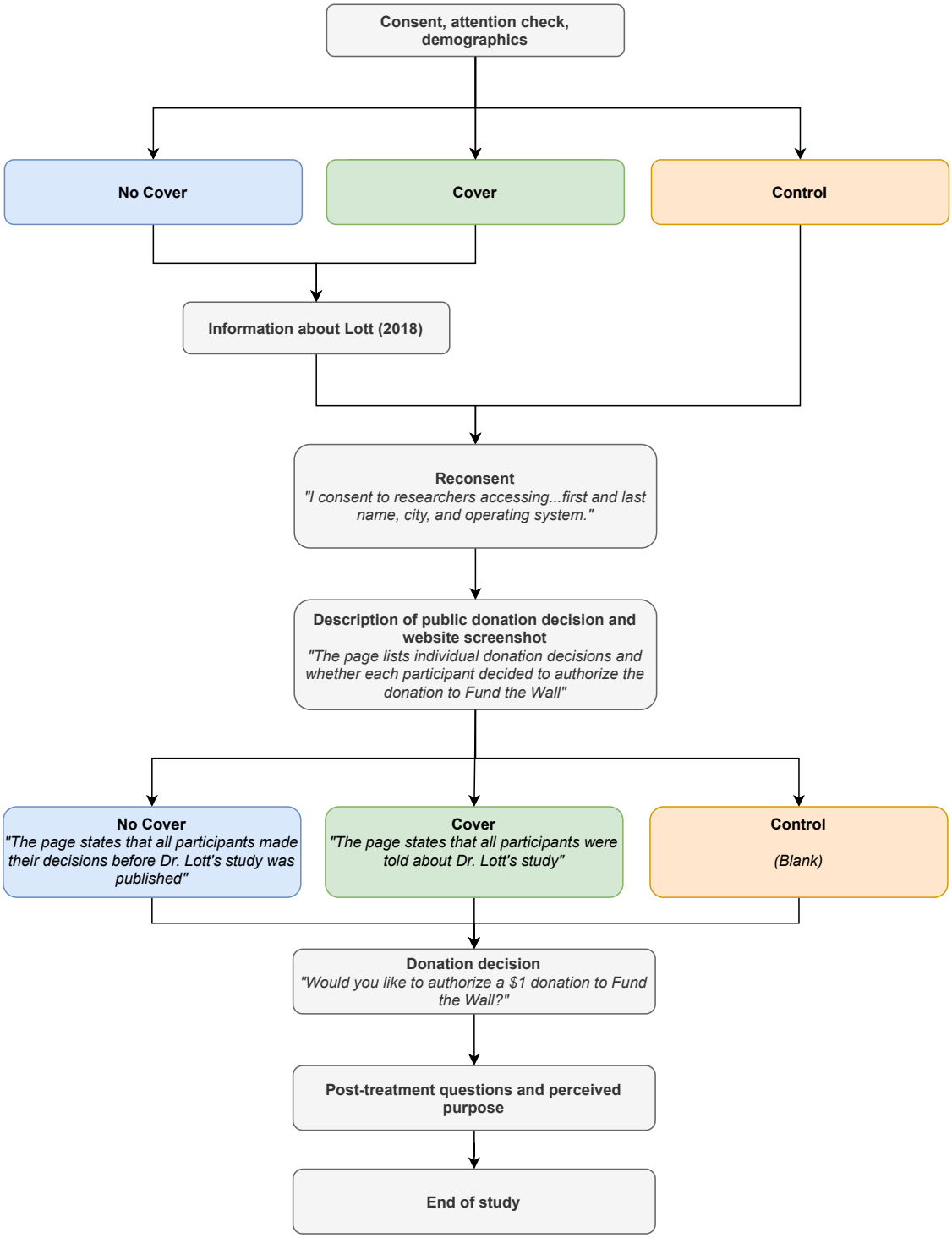
Visible donation decisions We ask respondents to consent to us accessing their name, city, and operating system from the survey provider (which confirmed that they would provide us with this data subject to participant consent) and give respondents the option to terminate the survey if they do not consent. Respondents are informed that they will have the opportunity to authorize a $1 donation to Fund The Wall, an organization seeking to construct the proposed US–Mexico border wall, and that we will post the results from the survey, including their individual donation decision, on our study website.[37] To vary the availability of a social cover, we tell respondents assigned to the *No Cover* treatment that the web page will state that "all participants were surveyed before Dr. Lott's study was published in an academic journal." In the *Cover* treatment, respondents are instead told that the web page will state that "all participants were shown the the preliminary findings from Dr. Lott's study before deciding whether or not to donate to Fund The Wall."

We also inform our respondents that "As researchers, we believe it is important to communicate our findings about political and social attitudes in [City of respondent] to the public".[38] We then inform our respondents that "We will promote our website via Facebook ads to [City of respondent] residents". This generates a plausible social cost for acting in a way that will be stigmatized in the respondent's area. After informing respondents about the content of the website, we ask people whether they "would like to authorize a $ donation to **Fund The Wall**?"

[37] To minimize experimenter demand concerns, we in fact informed our respondents that we would randomly select one of two organizations — *Fund the Wall*, or the *Texas Civil Rights Project*, an organization that (among other activities) worked to legally challenge the wall's construction — and that they would have the opportunity to authorize a $1 donation to this organization. In practice, we randomized almost all respondents to Fund the Wall to maximize statistical power for our comparison of interest.

[38] We used participants' IP address to capture and display their current location (i.e. their city).

Figure C2: Auxiliary Experiment 7: Design



C.4.2 Results

Average treatment effects To identify the joint effects of direct persuasion and anticipated persuasion of the audience (i.e. the direct persuasive effect of learning about the Lott study in addition to the indirect effect of learning that one's audience has learned about the Lott study and may thus be more likely to approve of the donation), we compare the *Control* condition with the *No Cover* condition. To identify the cover effect of rationales, we compare the *No Cover* condition to the *Cover* condition. This design thus allows us to benchmark the cover effect of rationales effect against the joint effect of persuasion and anticipated persuasion.

Table C4 displays mean donation rates by condition. We find a statistically significant effect on respondents' willingness to authorize a donation to Fund the Wall: respondents in the *Cover* condition are 7 percentage points more likely to authorize the donation than respondents in the *No Cover* condition (as shown in Column 6 of Table C4 which pools observations across all conditions). As shown in Panel A of Table C4, effect sizes are almost identical in our pre-specified main study, our pilot study, and a replication several months later. In contrast to the *Cover* vs. *No Cover* comparison, respondents in the *No Cover* condition are only 1 percentage point more likely to authorize a donation than respondents in the *Control* condition, suggesting that the combined effects of persuasion and anticipated persuasion are small. Relatively small persuasion effects are in line with other information provision experiments in the immigration domain, which typically find relatively small or null effects on behavior and stated preferences (Alesina et al., 2019; Hopkins et al., 2019; Grigorieff et al., 2020; Haaland and Roth, 2020), and are also consistent with the mixed evidence on persuasion we find in Experiments 3 and 4.

Heterogeneity by local vote shares The audience's composition — in this case, the fraction who approve of the decision to donate — should affect the magnitude of the rationale effect. Because we informed respondents that we would promote the website within their geographical area, we might expect that, controlling for respondents' own private views, respondents in areas with a greater fraction of Republicans (who are likely to approve of the decision to donate to Fund the Wall even in the absence of a rationale) should be less sensitive to the availability of a rationale than those in areas with a lower fraction of Republicans. We thus pre-registered investigating heterogeneity by the 2016 Republican vote share of the respondents' county, which we do by interacting our treatment indicators with vote shares (standardized, for ease of interpretation). We flexibly control for differential effect of partisanship by also controlling for the interactions between partisanship and the 2016 Republican vote share. Panel B of Table C4 displays the results, revealing striking heterogeneity by the Republican vote share of respondents' counties. The estimated interaction is large in magnitude: a one standard deviation increase in the Republican vote share of a respondent's county is associated with halving the magnitude of the cover effect of rationales. Of course, these results cannot be interpreted as a causal effect of differences in the composition of respondents' audiences: it may be, for example, that Republicans in Democratic areas feel a greater need to signal their support for the study by publicly donating.

C.4.3 Addressing experimenter demand concerns

One concern is that our instructions in Experiment 7, by directly informing (and reminding) respondents about their audience's information sets, induced experimenter demand effects or otherwise compromised external validity. *A priori*, it seems plausible that showing participants information

Table C4: Willingness to donate to anti-immigrant cause

	Dependent variable:					
	Donated to Fund the Wall					
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A:	*Non-interacted specification*					
Cover	0.063***	0.065***	0.074***	0.066**	0.053**	0.070***
	(0.020)	(0.018)	(0.016)	(0.028)	(0.026)	(0.014)
Control	−0.001	−0.005	−0.001			0.018
	(0.020)	(0.018)	(0.017)			(0.016)
p-value (Cover = Control)	0.0013	< 0.001	< 0.001			0.0012
Panel B:	*Interacted specification*					
Cover	0.061***	0.063***	0.071***	0.042	0.026	0.063***
	(0.020)	(0.019)	(0.016)	(0.030)	(0.028)	(0.014)
Cover × Rep vote share	−0.030	−0.038**	−0.038**	−0.062**	−0.073***	−0.042***
	(0.020)	(0.019)	(0.017)	(0.028)	(0.027)	(0.014)
Control	−0.001	−0.004	−0.001			0.017
	(0.020)	(0.018)	(0.017)			(0.016)
Control × Rep vote share	0.013	0.010	0.010			0.015
	(0.020)	(0.018)	(0.017)			(0.016)
Rep vote share	0.052***	0.0002	−0.0004	0.020	0.019	0.001
	(0.015)	(0.018)	(0.016)	(0.019)	(0.035)	(0.015)
Waves included	1	1	P + 1	2	2	P + 1 + 2
DV mean	0.488	0.488	0.497	0.484	0.484	0.494
DV std. dev.	0.500	0.500	0.500	0.500	0.500	0.500
Observations	3,751	3,751	4,457	1,279	1,279	5,736
Individual controls	No	Yes	Yes	No	Yes	Yes

Notes: The dependent variable is an indicator taking value 1 if the respondent donated to Fund the Wall. Columns 1–2 report results estimated on the sample from the Wave 1; Column 3 pools the sample from the Wave 1 with the sample from the pilot; Columns 4–5 consider only the sample from the Wave 2; and Column 6 pools all waves. In Panel B, the county Republican vote share is from the 2016 US Presidential election and is scaled to a standard normal distribution. The specifications in Columns 2–3 and Columns 5–6 of Panel B include the interactions between our set of partisan controls and the 2016 Republican vote share. Demographic controls include age, age squared, a set of race indicators, a Hispanic indicator, a male indicator, a set of education indicators. Partisan controls include indicators for "Very conservative", "Conservative", "Neither liberal nor conservative" (omitted), "Liberal", and "Very liberal". Robust standard errors are reported.

about the Lott study might induce demand effects and thus affect donation rates, but even if such demand effects are present, they do not bias our main comparison of interest (*Cover* vs. *No Cover*), given that participants in both treatment arms are shown identical information about the study. More concerning is the possibility the treatment manipulation of beliefs about the audience's information set induced differential experimenter demand effects between these two conditions.

Hand-coding of perceived purpose We also ask respondents to respond in open-ended form to the question "If you had to guess, what would you say the purpose of this study was?" We use responses to assess demand effects in two ways. First, we hired two independent research assistants to hand-code the responses. Appendix Table C5 in the Appendix shows that the majority of our respondents believed that we wanted to study the effects of information on anti-immigrant sentiment or participant's willingness to have their decisions posted on the website. Fewer than 1 percent of our sample correctly guessed the true purpose of our experiment (Column 1). The table also shows that on almost all of the dimensions we code, beliefs about the purpose of the study do not significantly differ between the *Cover* and *No Cover* conditions. The exception is Public Image (Column 3): respondents in the *Cover* condition are 2 percentage points more likely than respondents in the *No Cover* condition to believe that the study was about whether people were willing to publicly express political views. Although statistically significant, this difference is small in magnitude and cannot explain our effect sizes. We do find significant differences in perceived purpose between the *Control* condition the other two conditions, likely because we provided respondents in the *No Cover* and *Cover* conditions, but not in the *Control* condition, information about the Lott study. However, these differences do not affect our main comparison of interest (*No Cover* vs. *Cover*).

Table C5: Anti-immigrant donations: Perceived purpose of study

	Dependent variable:					
	Cover	Immigration attitudes	Public image	Information	Persuasion	Biased
	(1)	(2)	(3)	(4)	(5)	(6)
Cover	−0.005	0.010	0.019**	0.011	−0.012	−0.0004
	(0.003)	(0.015)	(0.010)	(0.015)	(0.012)	(0.014)
Control	−0.003	0.131***	0.037***	−0.013	−0.080***	−0.041***
	(0.003)	(0.016)	(0.010)	(0.016)	(0.012)	(0.014)
p-value (Cover = Control)	0.62	< 0.001	0.081	0.13	< 0.001	0.0042
DV mean	0.007	0.227	0.084	0.239	0.121	0.176
DV std. dev.	0.084	0.419	0.277	0.426	0.327	0.381
Observations	4,454	4,454	4,454	4,453	4,454	4,452
Individual controls	Yes	Yes	Yes	Yes	Yes	Yes

Notes: The dependent variable in each column is an indicator taking value 1 if the respondent's perceived purpose of the study was coded to fall into the corresponding category. "Cover" takes value 1 if the respondent correctly inferred the study was about whether the ability to link one's donation decision to the Lott study would affect the donation decision. "Immigration attitudes" takes value 1 if the respondent stated the study was about attitudes toward immigration. "Public image" takes value 1 if the respondent stated the study was about whether knowing one's decision will be observable to others would affect the donation decision. "Information" takes value 1 if the respondent stated the study was about disseminating information about immigration. "Persuasion" takes value 1 if the respondent stated the researchers were attempting to persuade them either to donate or not to donate. "Bias" takes value 1 if the respondent stated the researchers were biased. "Other" takes value 1 if the respondent stated a purpose that did not fall into any of the above categories. Categories other than "Other" are not mutually exclusive. All specifications pool the main experiment and the pilot and control for demographics and partisan affiliation. Demographic controls include age, age squared, a set of race indicators, a Hispanic indicator, a male indicator, a set of education indicators. Partisan controls include indicators for "Very conservative", "Conservative", "Neither liberal nor conservative" (omitted), "Liberal", and "Very liberal". Robust standard errors are reported.

Natural language processing Second, we use natural language processing techniques (in particular, BERT, introduced in Devlin et al. 2019) alongside a neural network classifier to predict treatment status given the participant's response to the open-ended elicitation of perceived purpose. The intuition is simple: if respondents do not perceive differences in the purpose of the study across conditions, there is no scope for experimenter demand effects to affect the results.

At a high level, our exercise proceeds in three steps. After splitting our sample into a training set (75%) and a test set (25%), we create high-dimensional vector representations, or contextual embeddings, of each respondent's answer to the question "If you had to guess, what would you say the purpose of this study was?". These embeddings capture semantic meaning.[39] We then train a neural network classifier on the training set to predict the respondent's treatment condition (*Control*, *Cover*, or *No Cover*) based on their contextual embedding. Finally, we calculate accuracy on the test set.

To generate contextual embeddings for each response, we use DistilBERT (Sanh et al., 2020), a transformer modeled on the BERT (Devlin et al., 2019) architecture that achieves similar performance at substantially lower computational cost. Our implementation is provided by HuggingFace; we use all default parameter values and train the model for the default three epochs.[40] This step outputs a 768-dimensional vector embedding for each respondent. We next train a neural network for sequence classification on the training set, again using the HuggingFace implementation with default parameter values, to predict the respondent's treatment condition based on their embedding. [41] Using this trained model, we predict treatment conditions in the test set and compare predicted conditions with actual conditions.

In order to facilitate comparing classifier accuracy when distinguishing between the *Cover* and *No Cover* condition vs. distinguishing between *Control* and the other two conditions, we repeat the exercise above three times, each time excluding from both the training set and the test set one of the three conditions. Our classifier achieves 51% accuracy when distinguishing between *Cover* and *No Cover*, and we are unable to rule out the null hypothesis of 50% accuracy ($p = 0.48$). In contrast, our classifier achieves 60% accuracy when distinguishing between *Control* and *No Cover* and 58% accuracy when distinguishing between *Control* and *Cover*, with both rates statistically distinguishable from 50% accuracy ($p < 0.001$). Thus, as expected, respondents hold different beliefs about the purpose of the study in *Control* vs. the other two conditions, but not between *Cover* and *No Cover*. Given the differences documented through the hand-coding exercise, we view this result as validation for our method, as it demonstrates that we would likely detect substantial differences in perceived purpose between the *Cover* and *No Cover* conditions if such differences were present.

C.4.4 Additional tables

[39] See Liu et al. (2020) for a review of contextual embeddings.

[40] See `https://huggingface.co/docs/transformers/model_doc/distilbert`.

[41] See `https://huggingface.co/docs/transformers/training`.

Table C6: Anti-immigrant donations: Balance of covariates

	Overall		Cover	No Cover	Control	p-values		
	mean	std.dev.	mean	mean	mean	(E=NE)	(E=C)	(NE=C)
Age	45.077	15.724	45.171	44.845	45.209	0.611	0.951	0.560
Black	0.076	0.265	0.070	0.089	0.069	0.075	0.965	0.063
Asian	0.043	0.203	0.041	0.042	0.045	0.886	0.588	0.690
White	0.824	0.381	0.828	0.815	0.830	0.414	0.887	0.330
Hispanic	0.106	0.308	0.112	0.105	0.102	0.580	0.419	0.804
Male	0.498	0.500	0.495	0.503	0.497	0.686	0.925	0.751
High school diploma	0.977	0.151	0.976	0.976	0.977	0.965	0.908	0.944
Bachelors degree	0.376	0.484	0.394	0.367	0.369	0.166	0.196	0.910
Republican	0.425	0.494	0.417	0.436	0.421	0.327	0.825	0.439

Notes: p-values based on robust standard errors reported. Attriters dropped from sample.

Table C7: Anti-immigrant donations: Sample representativeness

	Lott	Pew (Inds and Reps)
Age	45.11	47.17
Black	0.08	0.05
White	0.82	0.77
Asian	0.04	0.03
Hispanic	0.11	0.11
Male	0.49	0.52
High school diploma	0.98	0.93
Bachelors degree or higher	0.37	0.31
Observation	4553	5501

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.

C.5 Auxiliary Experiment 8: Interpretation of Lott Rationale

This section presents a pre-registered experiment in which we explore how learning that a previous participant had read the Lott study prior to making their donation decision shapes an audience's inference about the participant's motives and resulting social sanctions.[42] The experimental design is broadly similar to that of Experiments 2 and 4, but in addition to examining inference only about the matched participant's prejudice, we also examine inference about persuadability (or gullibility) — i.e. the extent to which a matched participant might be persuaded by a rationale. This is meant to capture arguments of the form "I have nothing against immigrants... I just believe the study." Of course, persuadability is only one of a set of potential reasons for donating after being exposed to information suggesting immigrants commit more crimes (which is why we choose not to explicitly study it in the main experiments presented in this paper); alternative reasons include lower tolerance for crime, higher levels of risk aversion, etc. We chose to focus on persuadability in this experiment because it is arguably the most natural "second type," because it was the most frequent reason cited in our pilot results, and because it is most consistently and objectively coded.

C.5.1 Sample and Experimental Design

Sample composition We recruited a sample of 3,047 Democrats through Luc.id in February 2020.[43] Our sample of respondents is broadly representative of Democrats and Independents in the United States (Table C8), though slightly more white and educated, and well-balanced on observables across treatment arms (Table C9).

[42] The pre-registration is available in the AEA RCT registry under ID AEARCTR-0005462.

[43] In our pre-registration, we specified that in some specifications, we would pool data from a pilot ($N = 2,019$) with the data from the main experiment. The pilot instrument was virtually identical to the instrument used in the main experiment. We report both unpooled and pooled specifications.

Table C8: Interpretation of Lott rationale: Sample representativeness

	Lott	Pew (Inds and Dems)
Age	41.35	45.86
Black	0.18	0.18
White	0.70	0.59
Asian	0.05	0.05
Hispanic	0.14	0.15
Male	0.45	0.46
High school diploma	0.98	0.89
Bachelors degree or higher	0.45	0.35
Observation	3133	6627

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 C9: Interpretation of Lott rationale: Balance of covariates

	Overall		Cover	No Cover	p-value
	mean	std.dev.	mean	mean	(R=NR)
Age	41.376	15.639	41.703	41.048	0.247
Black	0.182	0.386	0.186	0.179	0.612
Asian	0.045	0.208	0.049	0.042	0.386
White	0.710	0.454	0.703	0.716	0.455
Hispanic	0.140	0.347	0.136	0.144	0.561
Male	0.450	0.498	0.451	0.448	0.840
High school diploma	0.983	0.130	0.983	0.983	0.998
Bachelors degree	0.446	0.497	0.454	0.439	0.391

Notes: p-values based on robust standard errors reported.

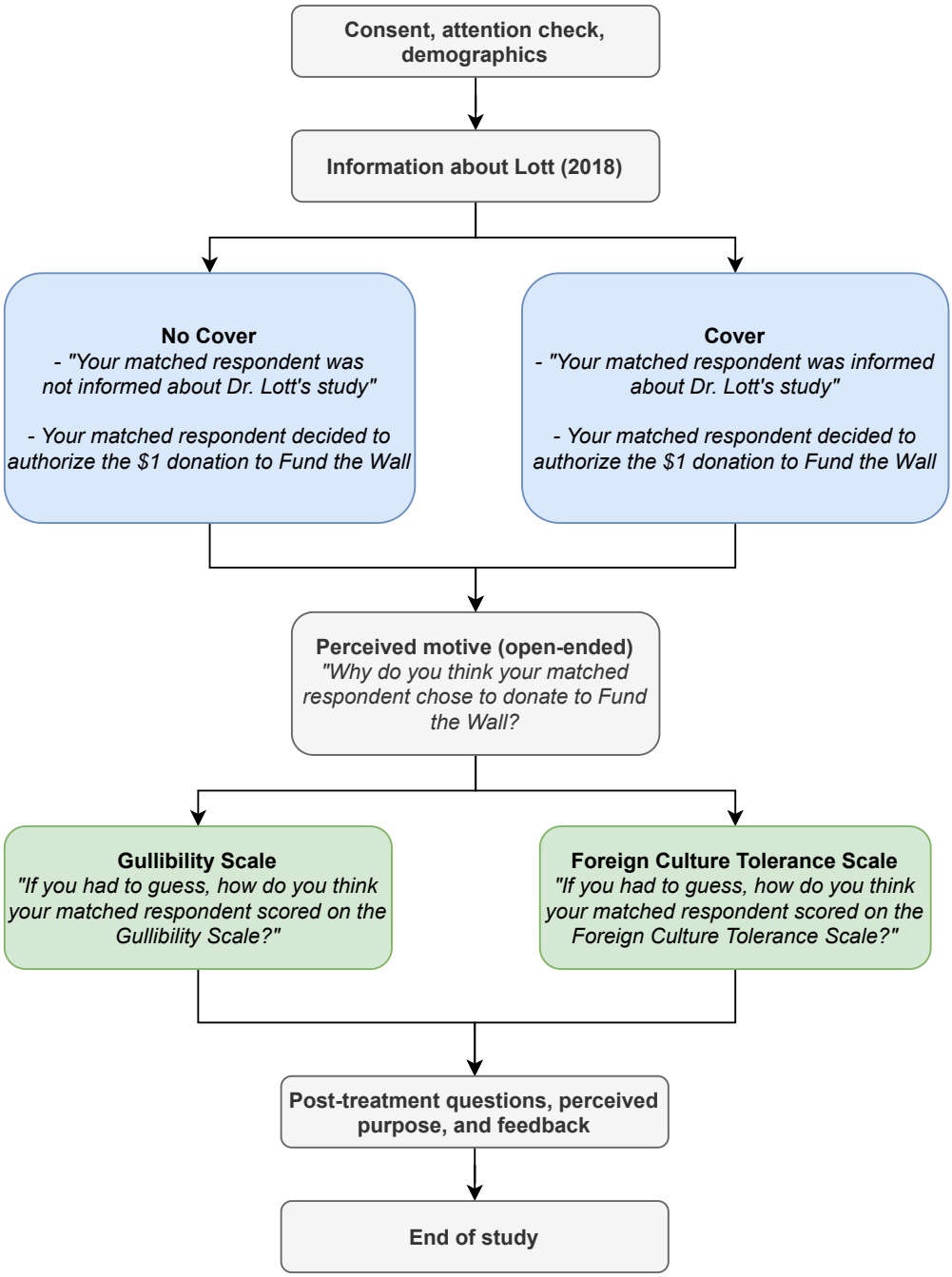
Information provision Figure C3 outlines the structure of the experiment. We tell all respondents about the Lott study (Lott, 2018) described in Section C.4 and about the fact that the study has been challenged on methodological grounds.[44] We then tell respondents that we conducted a project on political and social attitudes in the United States earlier in the year, and that participants in this previous study were given an opportunity to authorize a $1 donation to Fund the

[44]Once again, in order to ensure that our respondents are not misinformed, we debrief them at the end of the study and provide them with a meta-analysis summarizing the work on the effects of immigration on crime (Ousey and Kubrin, 2018).

Wall. We inform participants that we have matched them with one of these respondents, and that this respondent chose to authorize the donation. Respondents in the *Cover* condition are told that their matched respondent was informed about the study before deciding whether or not to authorize the donation to Fund the Wall, while respondents in the *No Cover* condition are told that their matched respondent was not informed about the study before making their donation decision.[45]

[45]Because all participants in the donations experiment were informed of the study before making their donation decision, in order to avoid deception, we ran a small auxiliary version of this experiment in which some respondents were *not* informed. These are the participants with whom we match respondents in the experiment described in this section.

Figure C3: Interpretation of Lott rationale: design



Measuring inference After learning whether or not their matched respondent knew about the study, all participants respond to the following open-ended question: "Why do you think your matched respondent chose to donate to Fund the Wall?" These open-ended responses form the raw data for our first measure of inference, which we analyze using the following (pre-registered) procedure. We begin with five "seed words" for each of the two dimensions: for intolerance, we chose *racist*, *biased*, *xenophobic*, *intolerant*, and *prejudiced*; and for persuadability, we chose *convinced*, *persuaded*, *gullible*, *naive*, and *sucker*. We add all "most relevant" synonyms for these words, as classified by the website `www.thesaurus.com`. In order to capture different parts of speech, we stem all words in our list (e.g., xenophobic → xenophob, gullible → gullib), for a total of 23 intolerance-related stems and 30 persuadability-related stems (Gentzkow et al., 2019). We then define two indicator variables, one for the presence of an intolerance-related stem and another for the presence of the persuadability-related stem, and we estimate treatment effects on the probability that the respondent uses at least one word in each list.[46]

For our second measure of inference, participants are then cross-randomized into one of two conditions: "tolerance" and "persuadability".[47] Participants in the "tolerance" condition are told that their matched respondent completed the "Foreign Culture Tolerance Scale," a "short questionnaire measuring tolerance toward foreign values and traditions," before making their donation decisions. Participants in the "persuadability" condition are told that their matched respondent completed the "Gullibility Scale," a "short questionnaire which measures how easily people are manipulated by evidence from untrustworthy sources," before making their donation decisions. All participants are asked to guess their respondent's score; we incentivize this guess by informing them that if they correctly guess the score, they will be entered into a lottery for a $50 Amazon gift card.

C.6 Results

Columns 1–3 of Table C10 displays results for our text-based measure of inference. Participants in the *Cover* condition are 7 percentage points less likely to use a stem related to intolerance when describing their matched respondent's motive, compared to a mean of 17 percent among participants in the *No Cover* condition ($p < 0.001$). These same participants are also 3 percentage points more likely to use words related to persuadability ($p < 0.001$), relative to a mean of 7 percent among participants in the *No Cover* condition.[48] These effect sizes highlight that the availability of a rationale strongly changes people's inference about their matched respondent's motives. That the effect on intolerance is larger than the effect on persuadability is consistent with the fact that persuadability is only one of several possible "second types" to which respondents might be substituting. These results are stable to the inclusion of demographic and partisan controls.

Columns 4–6 display results from our structured belief measures. Respondents in the *Cover* condition rate their matched participant 0.13 standard deviations lower on the intolerance scale ($p < 0.001$) and 0.32 standard deviations higher on the gullibility scale ($p < 0.001$) than partic-

[46] Responses that contain both an intolerance-related stem and a persuadability-related stem will have both intolerance and persuadability indicators equal to one, whereas responses that contain neither type of stem will have both indicators equal to zero. Thus, our results are unbiased even if participants perceive a nonzero correlation between intolerance and persuadability.

[47] We measure type inference using a "between" design (in which each respondent is asked only about a single dimension) rather than a "within" design (in which respondents are asked about both dimensions). We employ a between design in order to minimize experimenter demand effects and to avoid order effects (Haaland et al., 2021).

[48] We were intentionally conservative when choosing stem words in order to minimize the rate of false positives.

ipants in the *No Cover* condition. As with the text analysis measure, effects are similar in the pilot and in the pre-registered main experiment, are robust to the inclusion of control variables, and are precisely estimated. Our two measures of type inference are also highly correlated: on average, a respondent who uses a word related to intolerance (persuadability) when describing the matched respondent's motive rates the matched respondent as around half a standard deviation more intolerant (persuadable) than a respondent who does not use such a word.

Table C10: Interpretation of Lott rationale: type inference

	Used word			Inferred score		
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A:	*Inference about intolerance*					
Cover	−0.070***	−0.068***	−0.068***	−0.134***	−0.133***	−0.152***
	(0.012)	(0.012)	(0.012)	(0.051)	(0.051)	(0.039)
Individual controls	No	Yes	Yes	No	Yes	Yes
Observations	3,047	3,047	3,047	1,524	1,524	2,532
DV mean	0.137	0.137	0.137	0.000	0.000	0.000
DV std. dev.	0.344	0.344	0.344	1.000	1.000	1.000
Panel B:	*Inference about persuadability*					
Cover	0.031***	0.032***	0.032***	0.321***	0.310***	0.317***
	(0.010)	(0.010)	(0.010)	(0.050)	(0.050)	(0.039)
Observations	3,047	3,047	3,047	1,523	1,523	2,533
DV mean	0.084	0.084	0.084	0.000	0.000	0.000
DV std. dev.	0.278	0.278	0.278	1.000	1.000	1.000

Notes: The dependent variable in Columns 1–3 is an indicator taking value 1 if the respondent used a key word when describing their matched partner's motive. The dependent variable in Columns 4–6 is the respondent's (standardized) guess about their matched partner's score on the test. Panel A reports results for inference about intolerance; Panel B reports results for inference about persuadability. Demographic controls include age, age squared, a set of race indicators, a Hispanic indicator, a male indicator, a set of education indicators. Partisan controls include indicators for "Very conservative", "Conservative", "Neither liberal nor conservative" (omitted), "Liberal", and "Very liberal". Robust standard errors are reported.

Taken together, our evidence suggests that when judging others' motives, people believe that those who donated with a rationale are more persuadable and less intolerant than those who donated without a rationale.

D 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. Of course, all experiments obtained approval from multiple Institutional Review Boards.

D.1 Considerations related to information provision (Experiments 3–4 and Lott experiment)

The information on the link between illegal immigration and violent crime we provide to respondents (the clip from *Tucker Carlson Tonight* in Experiments 3–4 and the Lott (2018) study in the robustness experiment) 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 either piece of evidence — indeed, we explicitly inform respondents in the Lott experiment that the study has been challenged by reputable sources — 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.[49]

***Tucker Carlson Tonight* clip** In Experiments 3–4, we provide respondents with video clip from *Tucker Carlson Tonight*, the most popular cable news show in the country. While we do not endorse the message, the raw numbers presented in the video clip are taken from the U.S. Sentencing Commission and are factually correct. While we debrief respondents with the above-described meta-study at the end of the survey, this is strictly speaking unnecessary, as the numbers cited in the video clip are not factually wrong.

Lott study In the Lott experiment, treated respondents receive information about a study by John Lott claiming that illegal immigrants commit more crime. Lott holds a PhD in economics from UCLA and has previously held positions at Yale University, the University of Chicago, and the Department of Justice. The working paper we describe — Lott (2018) — has been extensively cited by the Trump administration and is posted on the Social Science Research Network. In other words, it is a real study by an academic economist. The study has been challenged on methodological grounds, but we inform all respondents about this controversy on the very same screen where we present the study.[50] While we again debrief respondents, this is not strictly necessary, as respondents had already been informed that the study's methodology had been challenged.

[49] It is common practice to mislead respondents by omitting relevant details from the instructions. Indeed, correspondence studies generally rely on outright deception by sending out fake resumes to employers.

[50] Specifically, at the information page, all respondents are informed about the following: "However, a number of sources (including a researcher affiliated with the Cato Institute, a non-partisan libertarian think tank) have recently

D.2 Considerations related to privacy and deception (Experiments 1, 3, and Lott experiment)

Given that our mechanism examines the effect of perceived social stigma on behavior, it is crucial that respondents in Experiments 1 and 3 and in the Lott experiment believe that their decisions will be visible to others. Although our experiments avoid explicit deception, protecting participants' privacy in this context 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 D.3.

D.2.1 Experiments 1 and 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) and December 1, 2021 (Experiment 1). 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.

D.2.2 Lott experiment

Identifiable donations We asked participants to consent to us accessing their name and city from their survey provider (which confirmed that we could collect this data subject to participant consent). All participants had the option to terminate the survey if they did not consent. We informed those that consented that upon the publication of the Lott study in an academic journal, we would post the results from the survey, including their individual donation decision, on our study website. While we intend to do so should the study be published in an academic journal, this statement is somewhat misleading because it is unlikely that the study will ever be published (given its methodological errors and the fact that Lott has rarely published in peer-reviewed academic

challenged some of the study's methods, claiming that errors in analysis invalidate its results. Dr. Lott has responded to this critique with a defense of the study's methods, but the issue remains unresolved."

journals over the past decade). Despite participants consenting to us accessing and publishing their names, and the fact that only a small minority of the Republican and Independent participants are likely to be uniquely identifiable based on their first and last name alone (i.e. absent geographical or other identifiers), we still viewed it as desirable to preserve their anonymity: the formulation of our experiment allows us to do so with high probability.

Cover manipulation Conceptually, in the *No Cover* condition for the donations experiment, we would like to show respondents a website screenshot stating that "No participants were told about Dr. Lott's study." However, because these participants did in fact learn about the study, such a screenshot would be deceptive. Instead, we exploit the fact that Lott's study had not yet been published in an academic journal (a fact about which we explicitly informed all respondents when describing the website). In particular, we show respondents a website screenshot stating that "We surveyed respondents earlier this year before Dr. Lott's study was published." In the survey, we write that "the website states that you were surveyed before the study was published and does not mention that you were shown an early summary of the study's findings." Respondents in this condition thus believe that their audience will believe that they (the respondents) had no information excusing their decision to donate to Fund the Wall.

This formulation is misleading in that it relies on an academic, rather than commonplace, understanding of the word "published" (that is, "published in an academic journal" rather than "made available for public readership"). However, survey respondents themselves are not misled, as they are fully aware of the study's status and are fully aware of what others reviewing the donation decisions are likely to believe. The group that may be misled is thus the group who visit the website listing donation decisions. Given the low probability that this website will ever be published, we and our Institutional Review Board felt comfortable using this formulation.

D.3 Considerations related to subject pool contamination (Experiments 1, 3, and Lott experiment)

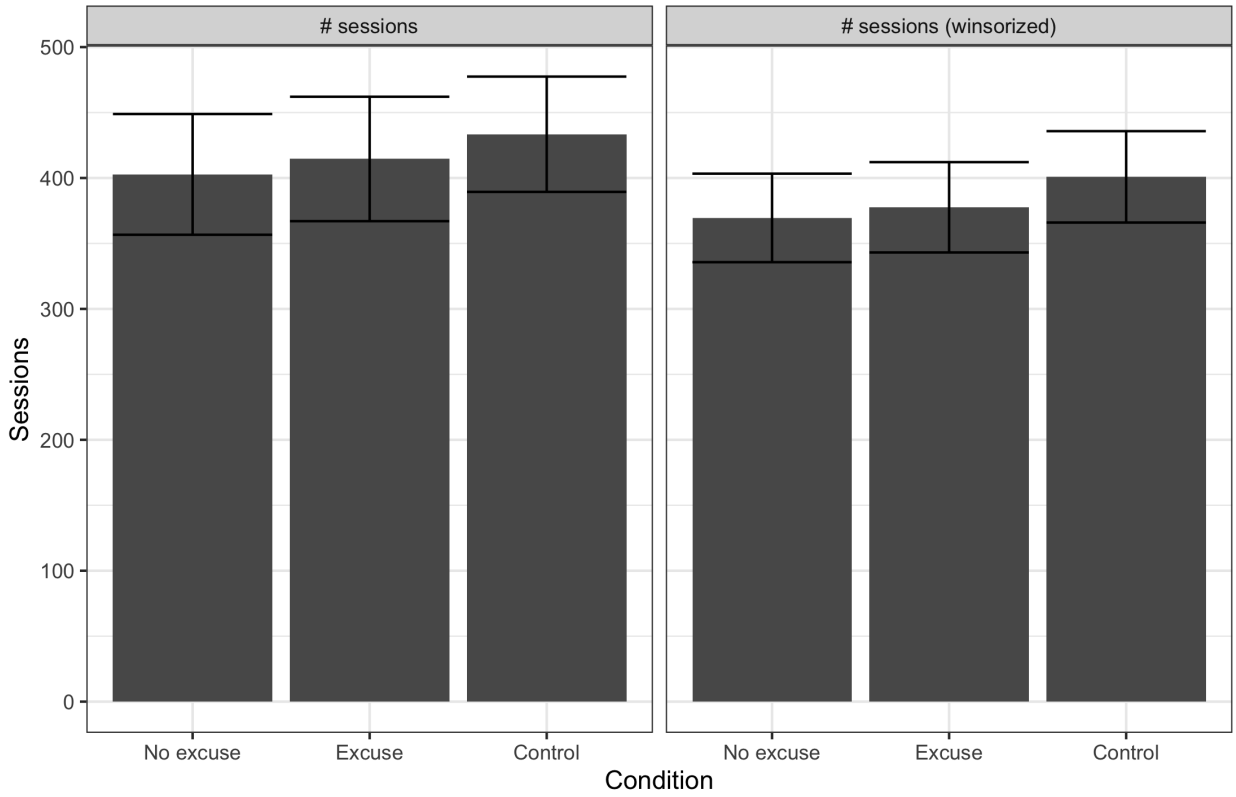
An important concern with deceptive or misleading experiments is that they can contaminate the subject pool by lowering trust in scientists and making respondents less likely to participate in future research studies. Of course, this can only happen if respondents know that they are being misled.

In the Lott experiment, subjects are told we will publish the website once Lott's study is published in an academic journal. Although we privately believe that the Lott (2018) study is very unlikely to be published in a journal, subjects do not know (and never learn) this. (If the study is published, we are prepared to set up the web page in line with the instructions provided to respondents.) Similarly, in Experiments 1 and 3, 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) or December 1, 2021 (Experiment 1). 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). To further substantiate the claim that our experiments had no effect on respondents' trust in social science experiments, we asked Luc.id to calculate the number of studies in which each respondent participated in the following nine months. In Figure D1, we examine

whether this number varies by experimental condition. We find no differences across the three treatment conditions (including the Control condition, in which respondents were not exposed to any information about immigrants).

While it is technically possible that the experimental conditions induced differential trust in social science that was not reflected in the number of studies in which respondents participated, we view this contingency as unlikely in light of the reasoning above. 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.

Figure D1: Subsequent survey behavior of respondents



Notes: Figure D1 presents the results of our analysis of the subsequent survey behavior of the respondents in the Lott experiment between the end of collection and December 2020. The figure presents the mean number of surveys (top panel) and the mean of the winsorized number of surveys (bottom panel) in which respondents participated, with the winsorization at the 0.98 quantile.

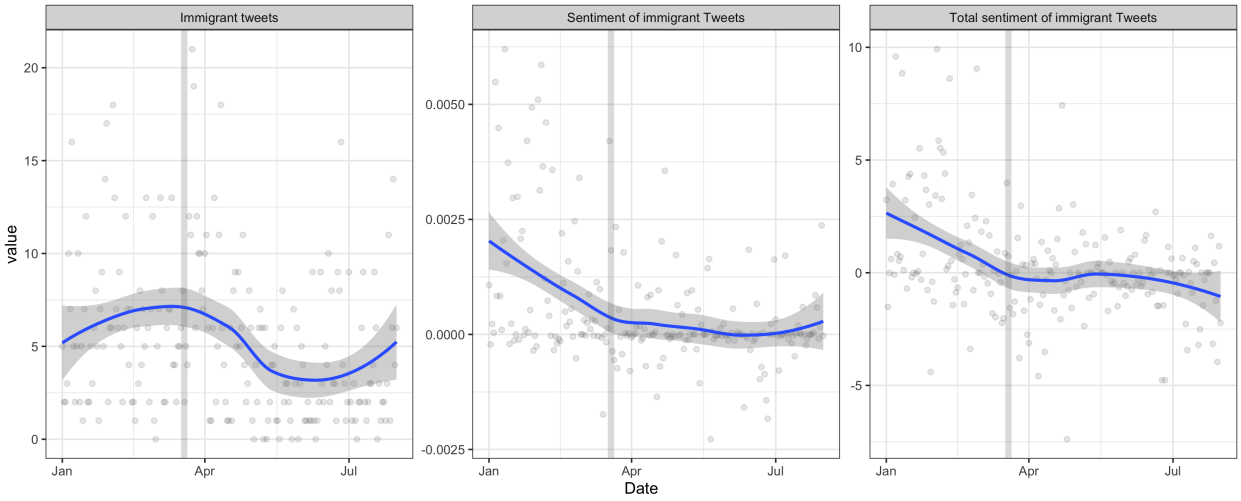
D.4 Considerations related to starting political Twitter campaigns (Experiments 1 and 3)

As discussed in Appendix D.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 D2 and Table D1. We find no evidence that respondents in our experiment begin posting more immigrant-related Tweets or more negative content about immigrants after participating (Figure D2). 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 D1). This evidence further strengthens our confidence that our experiment did not contribute to anti-immigrant discourse on social media.

Figure D2: Twitter activity of respondents before and after experiment



Notes: Figure D2 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 D1: Subsequent Twitter behavior of respondents

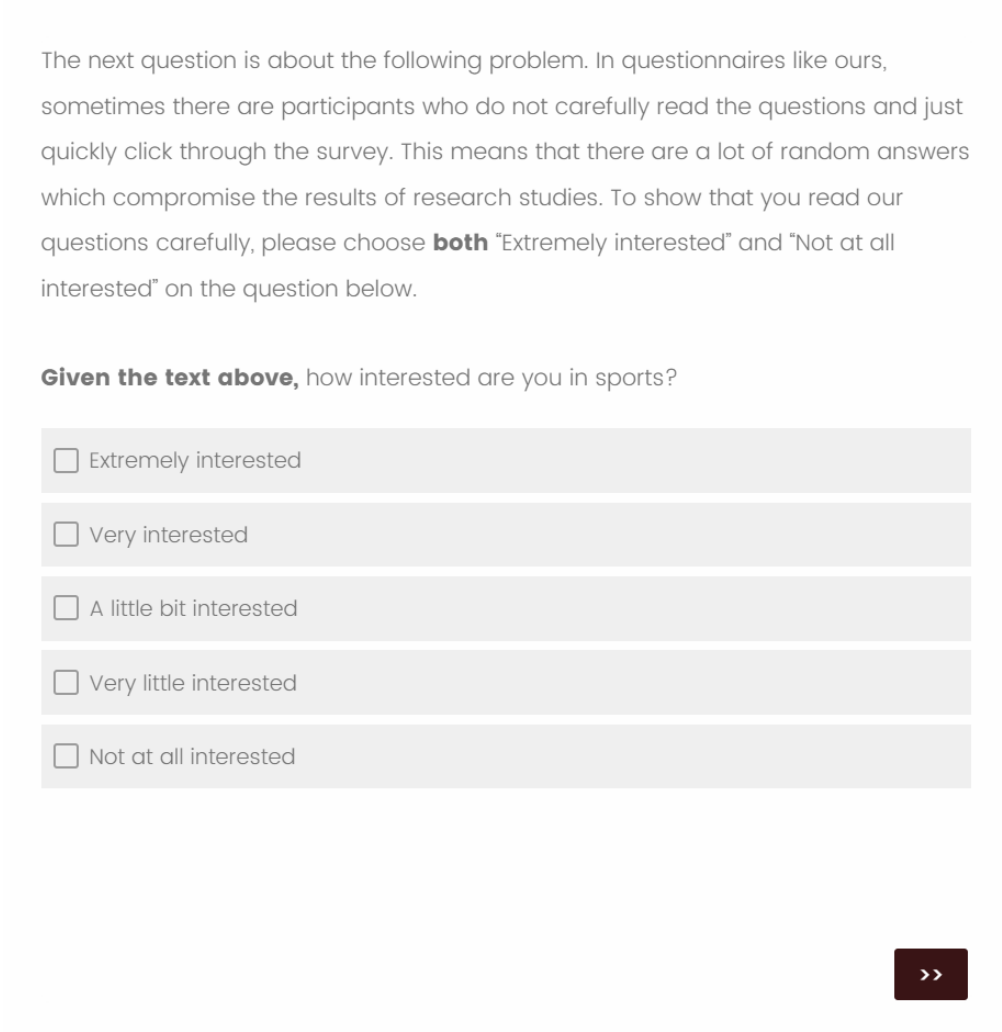
	Dependent variable:					
	Tw.	Tw. (w)	Imm. Tw.	Imm. Tw. (w)	Imm. sent.	Tot. imm. sent.
	(1)	(2)	(3)	(4)	(5)	(6)
Cover	−44.414	−9.298	−0.583	−0.152	0.005	0.024
	(29.941)	(9.462)	(0.416)	(0.117)	(0.012)	(0.062)
Constant	80.075***	35.951***	0.970***	0.383***	0.003	−0.052
	(20.862)	(6.593)	(0.290)	(0.082)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

Notes: Table D1 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.

E Experimental Instructions

E.1 Experiment 1: Expression of dissent – Democrats

E.1.1 Attention screener



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.

Sign in with Twitter

E.1.3 Background questions

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?

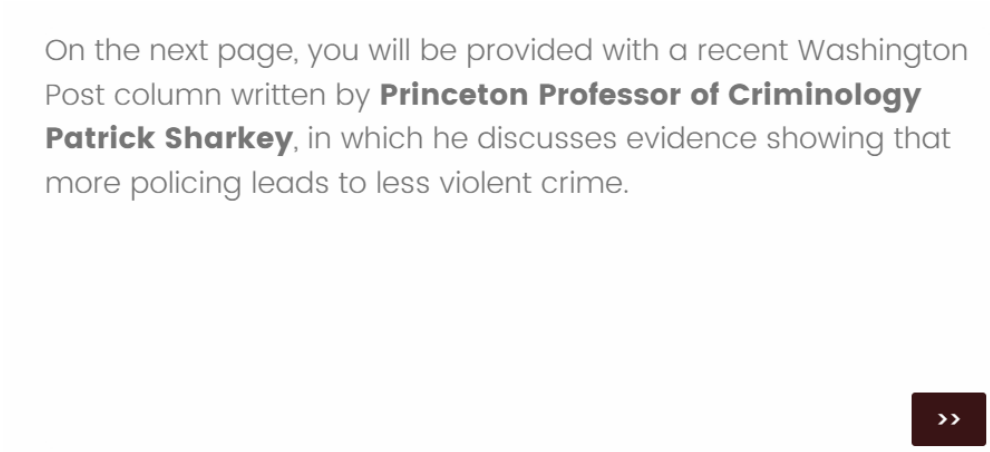
- ◯ Donald Trump
- ◯ Joe Biden
- ◯ Other
- ◯ Did not vote

Are you liberal or conservative?

- ◯ Very liberal
- ◯ Liberal
- ◯ Neither liberal nor conservative
- ◯ Conservative
- ◯ Very conservative

>>

E.1.4 Pre-treatment outcomes



Vincent Cecil for The Washington Post

Outlook

Why do we need the police?

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

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.

>>

Vincent Cecil for The Washington Post

Outlook

Why do we need the police?

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

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.)

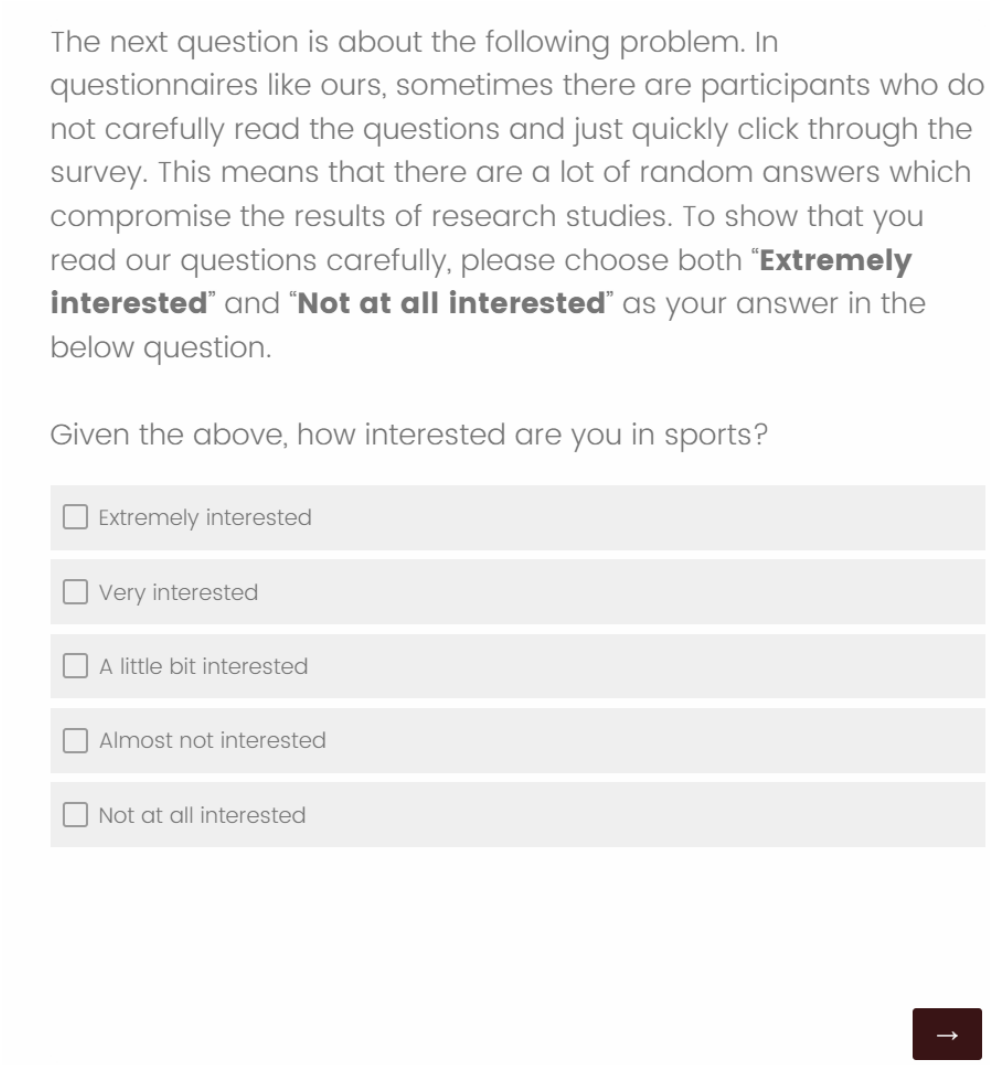
- ○ Yes
- ○ No

››

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

E.2 Experiment 2: Interpretation of dissent – Democrats

E.2.1 Attention screener and background questions



What is your sex?

- Male
- Female

What is your year of birth?

What is your political affiliation today: Republican, Democrat, or Independent?

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.

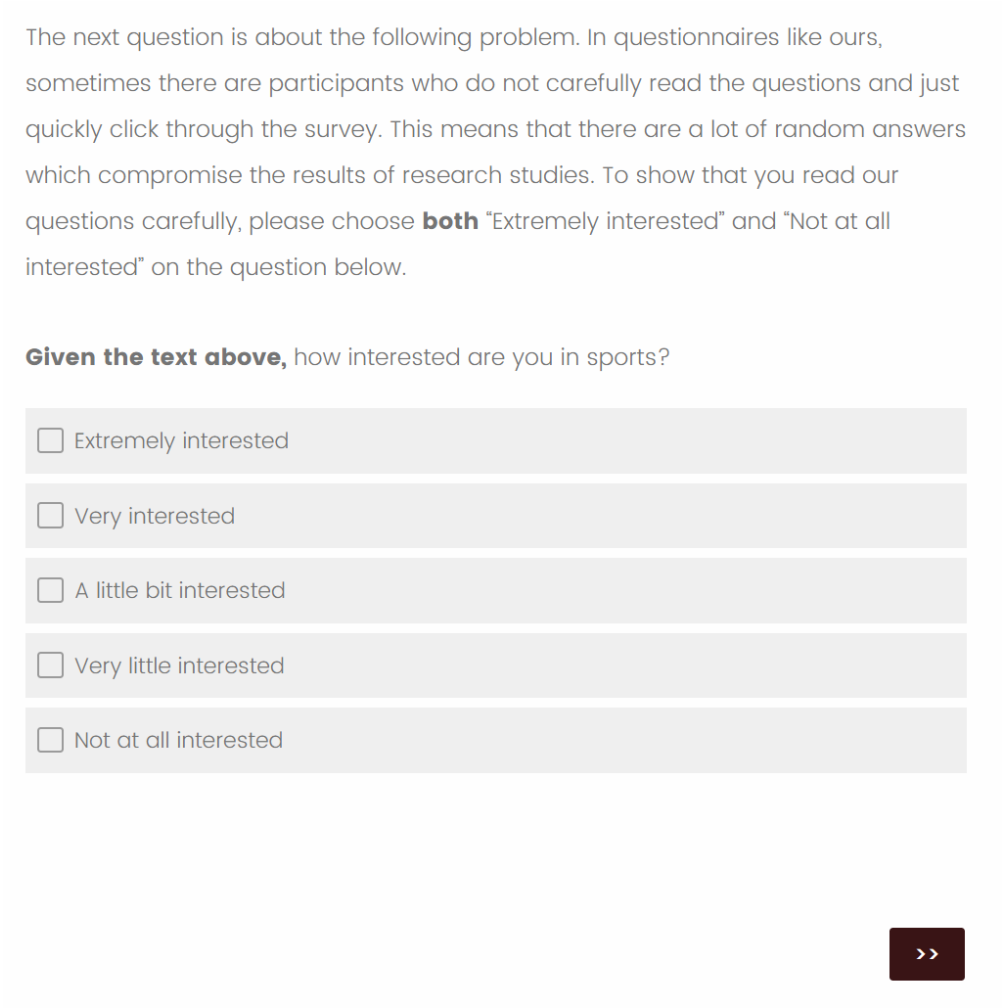
Do you want to authorize a $1 bonus to your matched respondent?

- ◯ Yes, I would like to authorize a $1 bonus
- ◯ No, I would not like to authorize a $1 bonus

→

E.3 Experiment 3: Expression of dissent – Republicans

E.3.1 Attention screener



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):

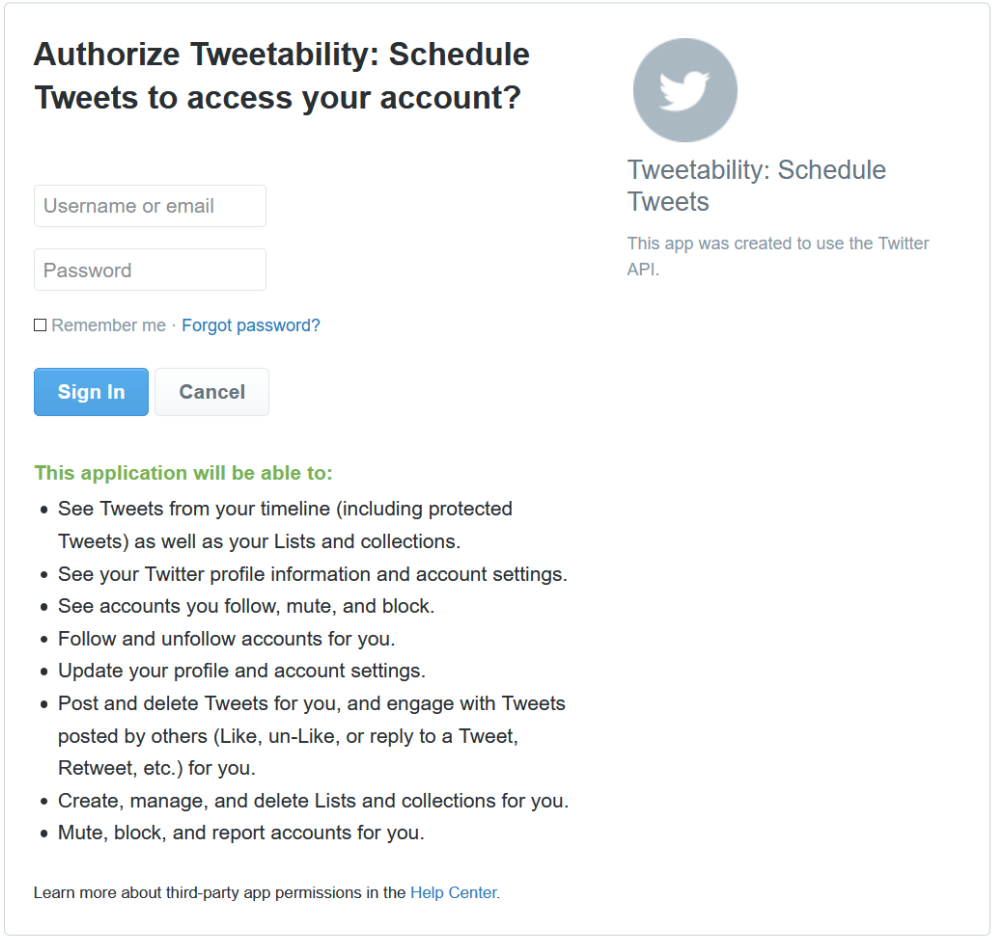
- We will **never** use the app to access non-public information from your account (including your posts)
- We will **never** use the app to make posts on your account without your **explicit consent**
- The app **does not give us access to your direct messages or email address**
- All identifying information 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

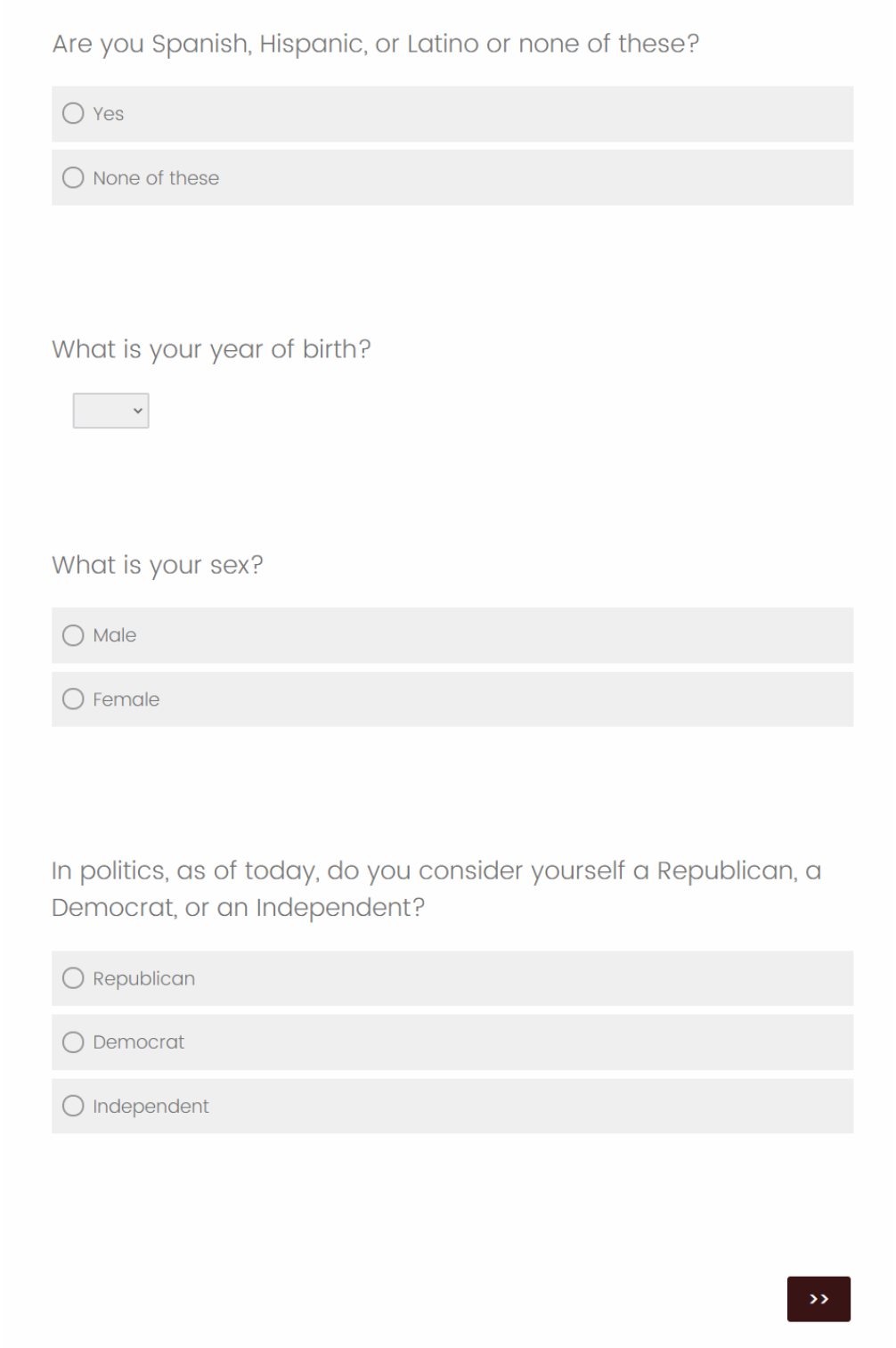
If you have any questions or complaints, you can contact the Social and Behavioral Sciences Institutional Review Board at the University of Chicago at:
The Social & Behavioral Sciences Institutional Review Board,
University of Chicago
Phone: (773) 834-7835
E-mail: sbs-irb@uchicago.edu

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.

Sign in with Twitter


Authorize Tweetability: Schedule Tweets to access your account?
Tweetability: Schedule Tweets
This app was created to use the Twitter API.
Username or email
Password
Remember me · Forgot password?
Sign In
Cancel
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



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?

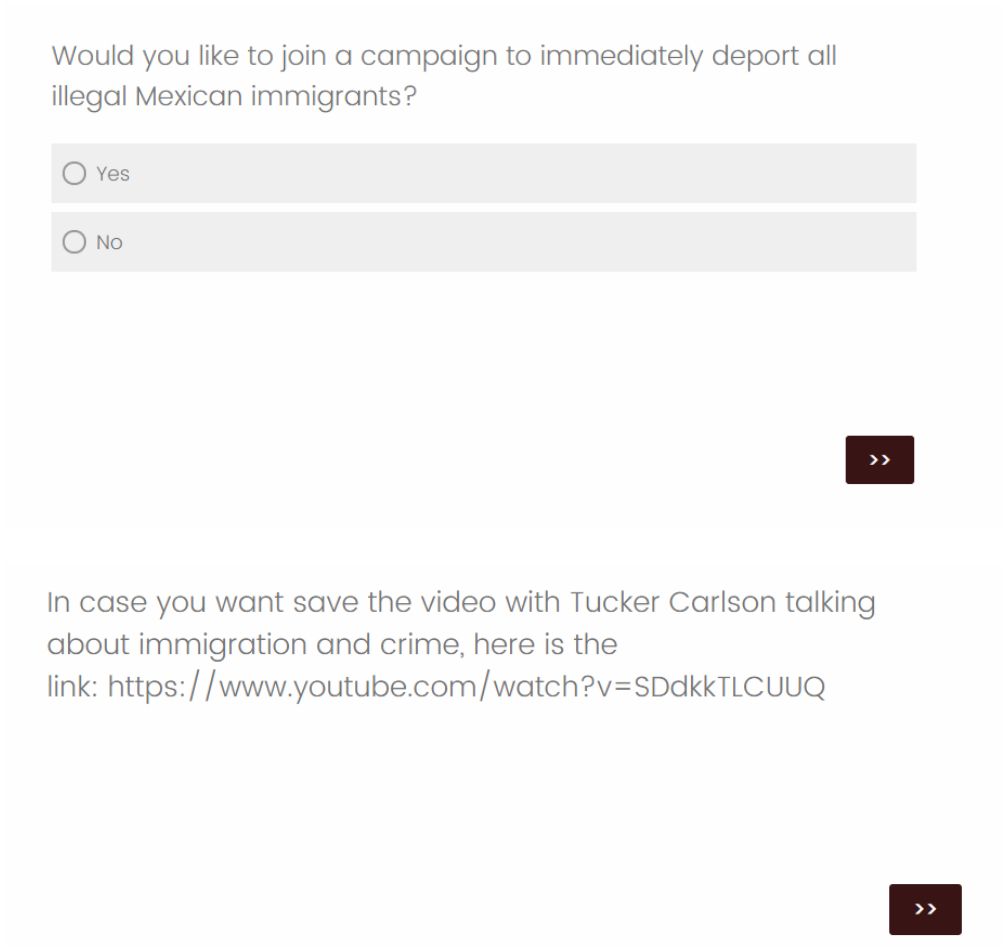
- ◯ Very liberal
- ◯ Liberal
- ◯ Neither liberal nor conservative
- ◯ Conservative
- ◯ Very conservative

>>

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



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

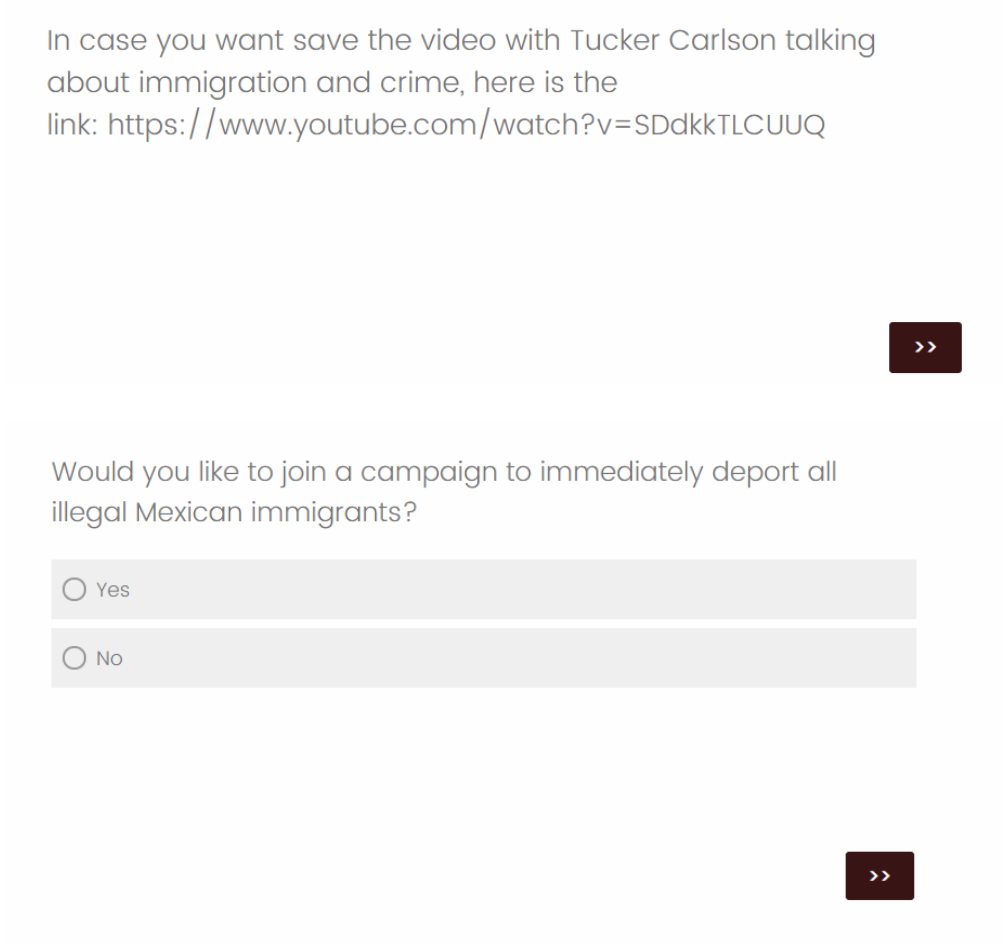
Tucker: Why didn't we know truth about illegals and crime?
Tucker's Thoughts: For years, we were told illegal immigrants were more law-abiding than American citizens. In fact, the ...
youtube.com

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



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

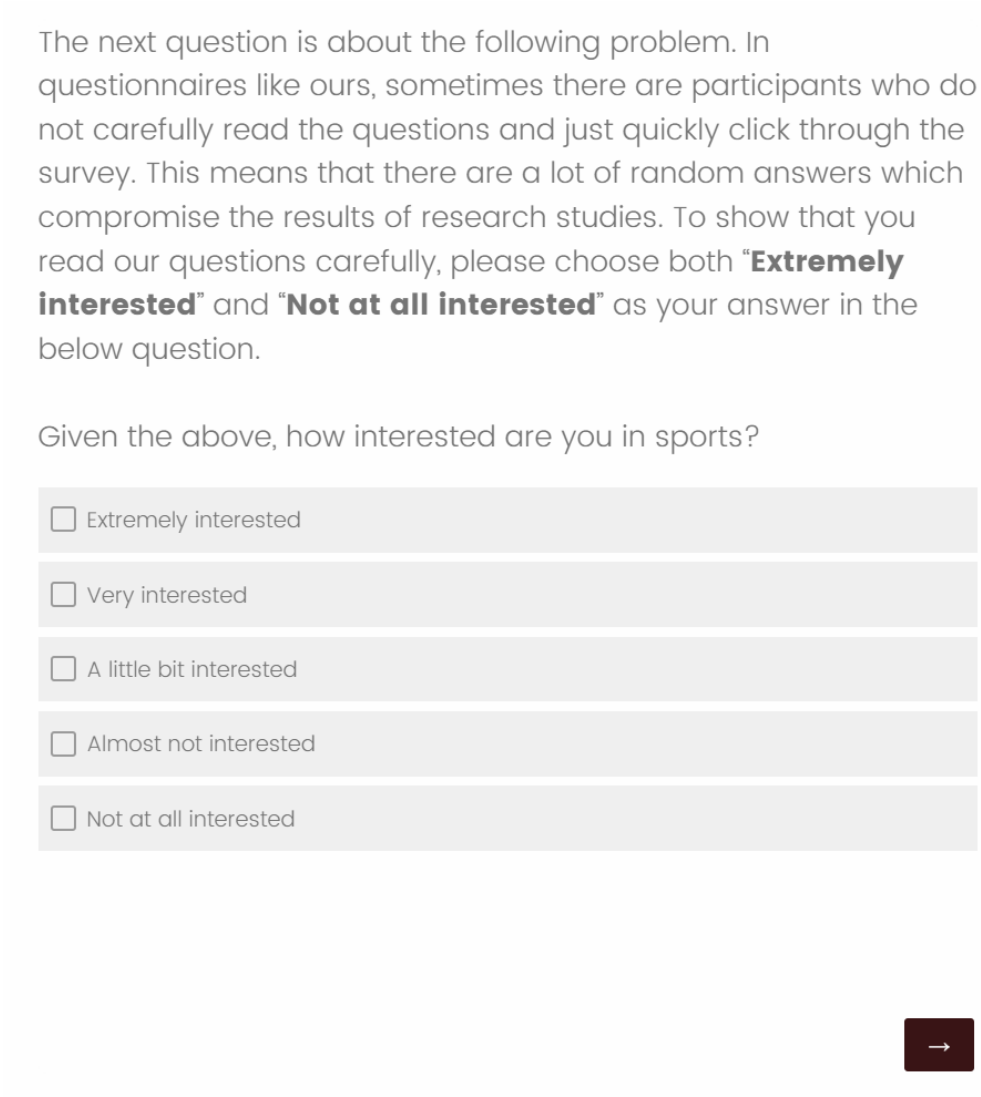
Do you authorize the *Tweetability* app to schedule the post above to be posted on your account? (If you choose "no," then nothing will be posted on your account.)

- ◯ Yes
- ◯ No

>>

E.4 Experiment 4: Interpretation of dissent – Republicans

E.4.1 Attention screener and background questions



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?

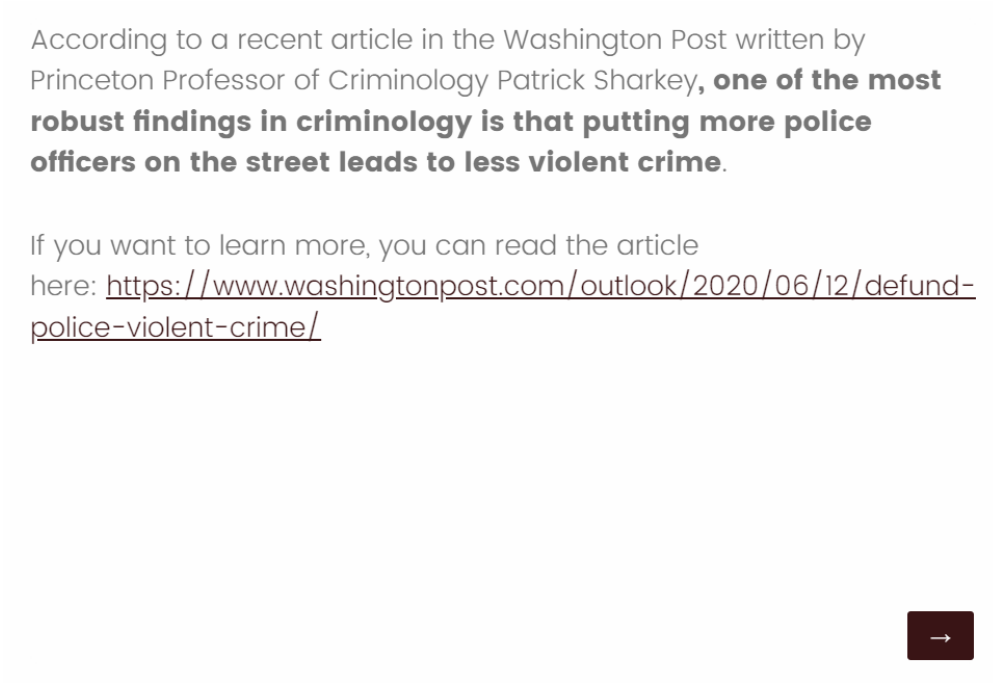
- ◯ Yes, I think my matched respondent chose to authorize the donation
- ◯ No, I think my matched respondent **did not** choose to authorize the donation

→

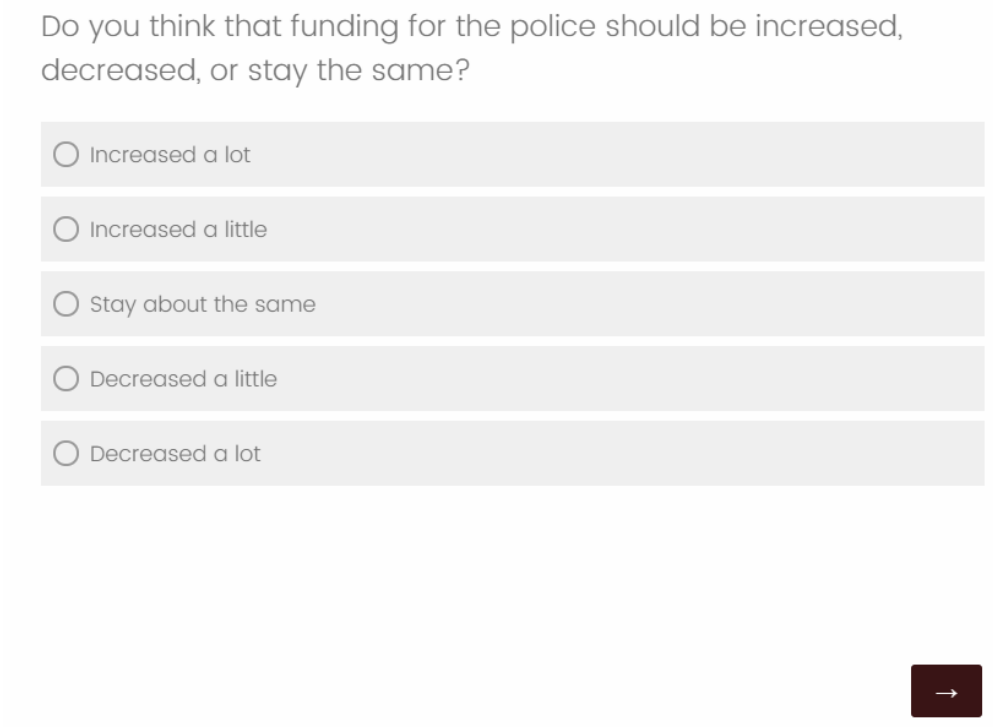
E.5 Auxiliary Experiment 1: Persuasion experiment – Democrats

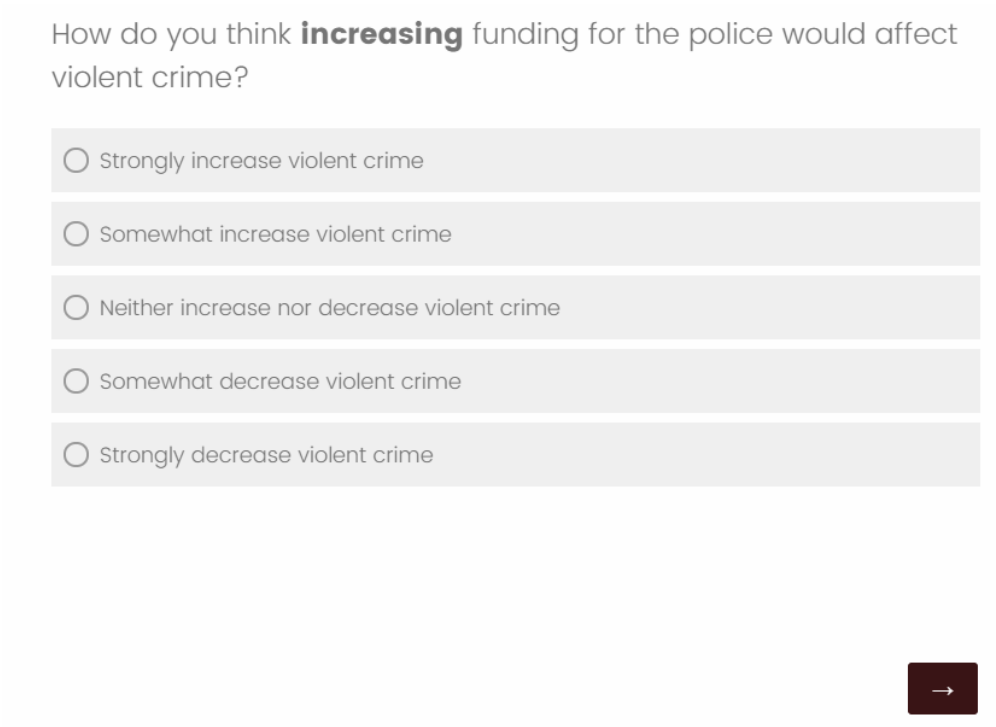
E.5.1 Pre-treatment beliefs

E.5.2 Information treatment (treatment group only)



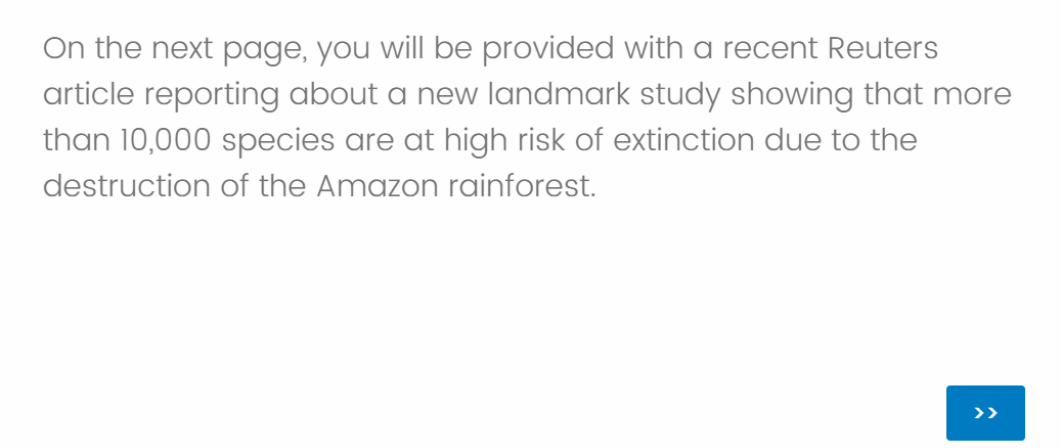
E.5.3 Post-treatment outcomes





E.6 Auxiliary Experiment 2: Rainforest placebo

E.6.1 Pre-treatment questions



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.6.2 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

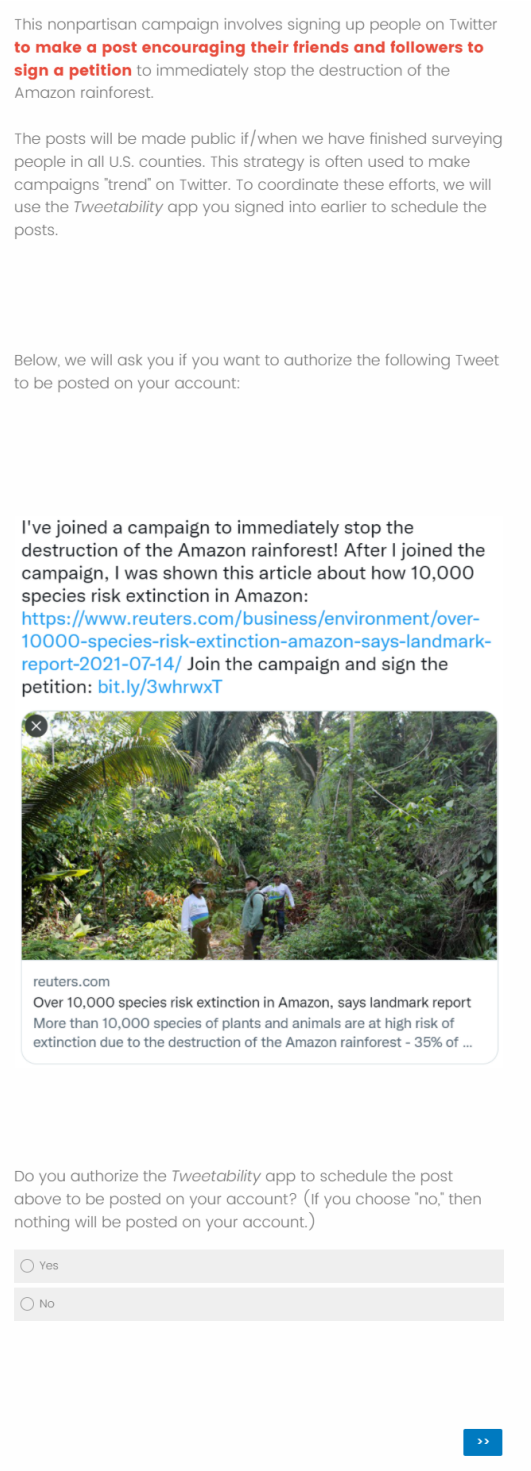
reuters.com
Over 10,000 species risk extinction in Amazon, says landmark report
More than 10,000 species of plants and animals are at high risk of extinction due to the destruction of the Amazon rainforest - 35% of ...

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

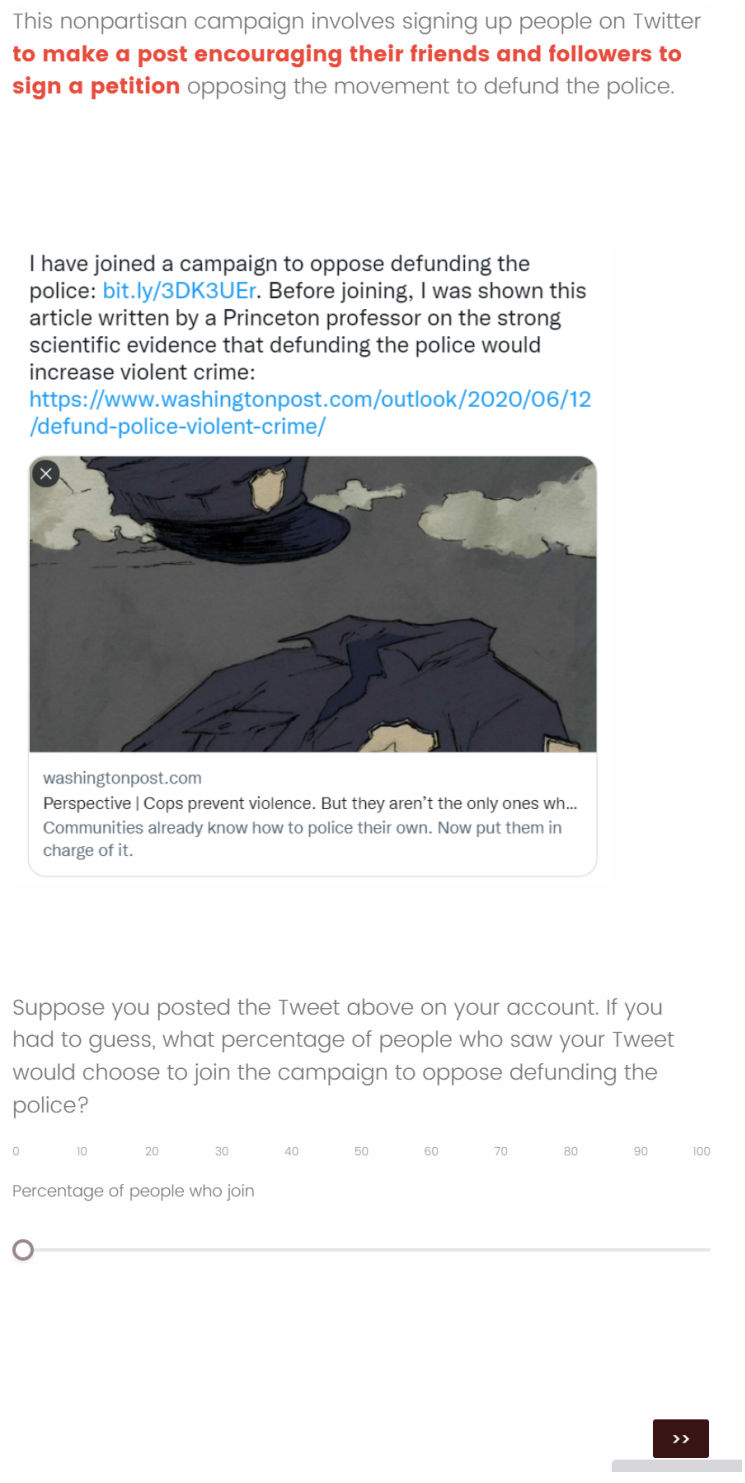
>>

E.6.3 Treatment: "After" wording (no rationale)

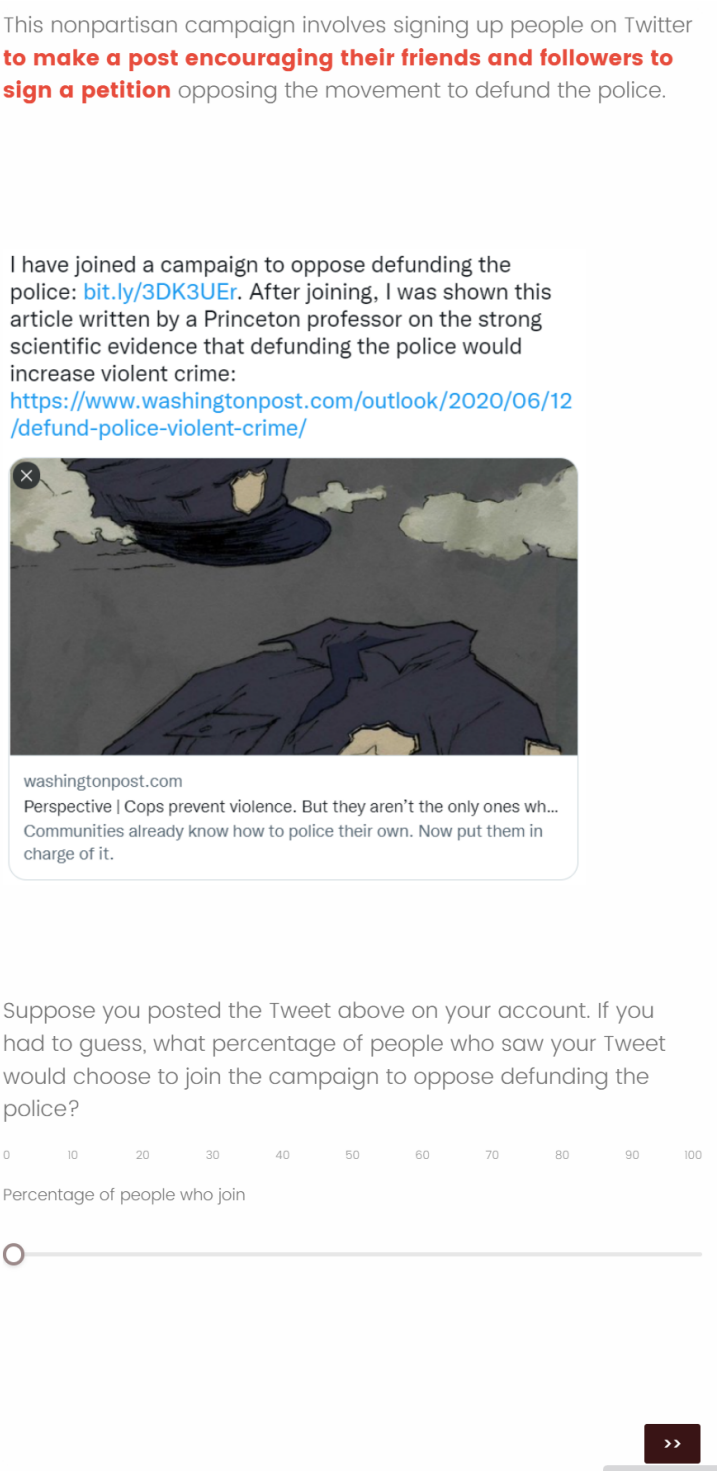


E.7 Auxiliary Experiment 3: Anticipated persuasion – Democrats

E.7.1 Treatment: "Before" wording (rationale)

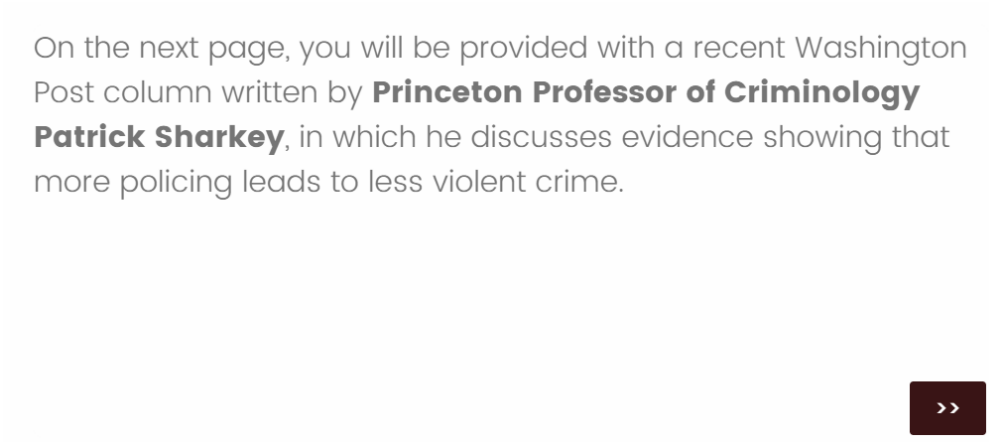


E.7.2 Treatment: "After" wording (no rationale)



E.8 Auxiliary Experiment 4: Open-ended explanations of preferred anti-defunding Tweet – Democrats

E.8.1 Pre-treatment questions



Vincent Cecil for The Washington Post

Outlook

Why do we need the police?

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

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.

Vincent Cecil for The Washington Post

Outlook

Why do we need the police?

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

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.8.2 Treatment: "Before" wording (rationale)

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

E.9 Auxiliary Experiments 5: Interpretation of dissent with low-credibility rationale

E.9.1 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.9.2 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

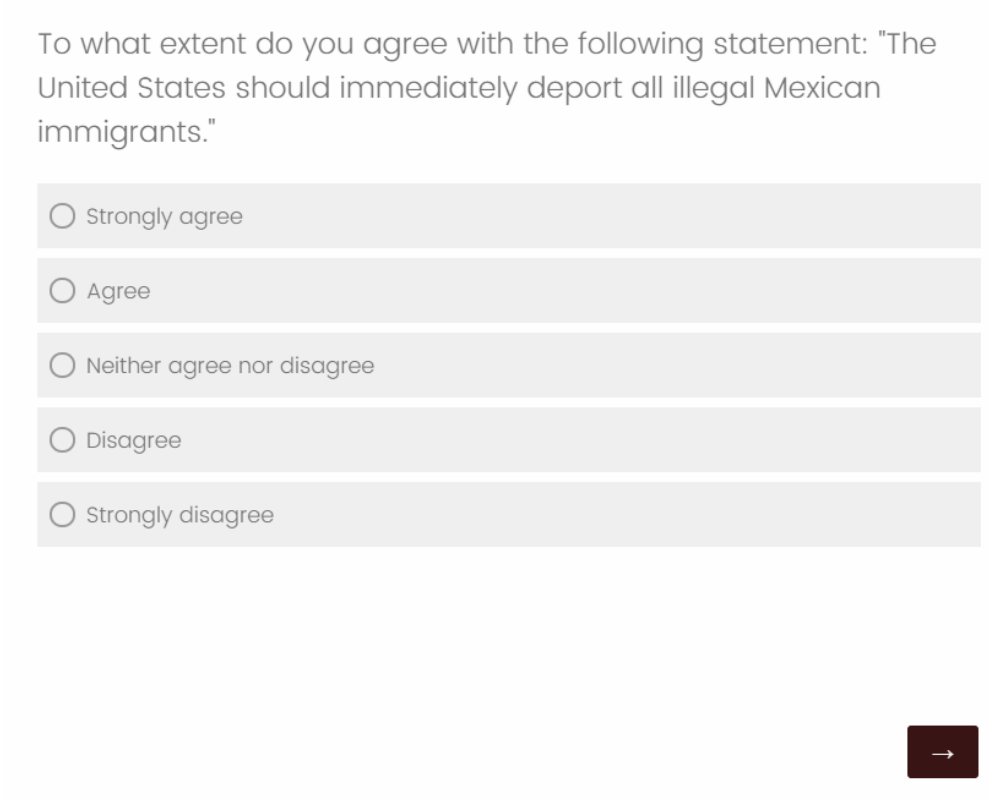
→

E.10 Auxiliary Experiment 6: Persuasion experiment – Republicans

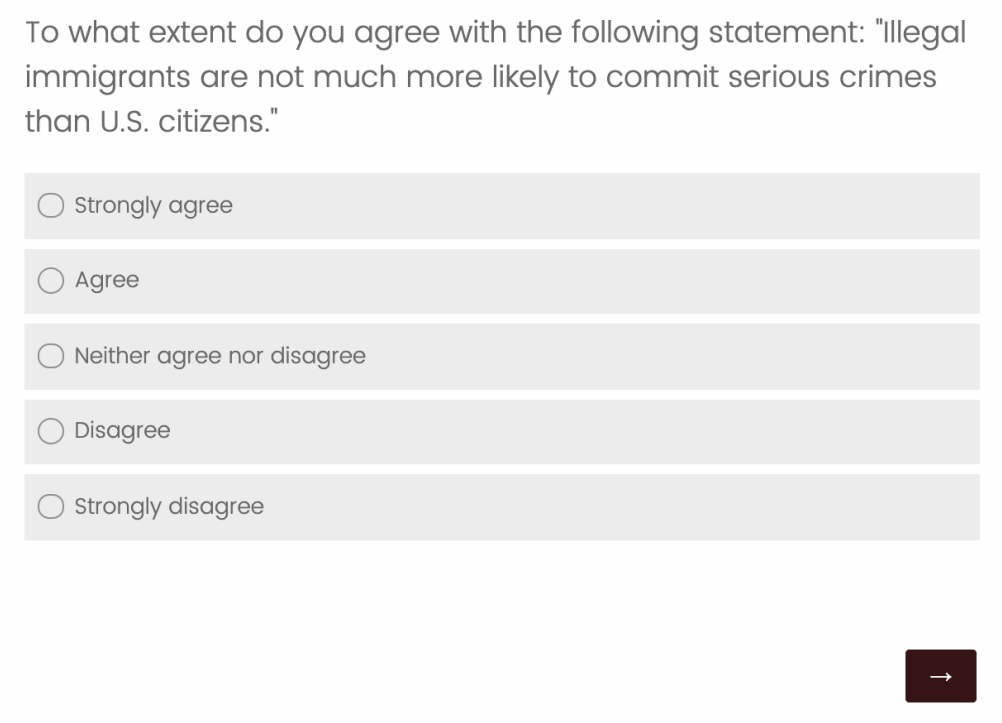
E.10.1 Pre-treatment beliefs

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

E.10.2 Information treatment (only shown to respondents in the treatment group)



E.10.3 Post-treatment outcomes



E.11 Auxiliary Experiment 7 (wave 2): Expression of dissent with visible donation

E.11.1 Pre-treatment questions

Do Illegal Immigrants Commit More Crime? Preliminary evidence from Dr. Lott's study

Economist Dr. John R. Lott carried out a study on the relationship between illegal immigration and crime. **The study has not been published in an academic journal, but we summarize some of its findings here.**

Dr. Lott's study uses **detailed data spanning over two decades from the Arizona state prison systems**. He finds that **"undocumented immigrants are at least 142% more likely** to be convicted of a crime than other Arizonans. They also tend to commit **more serious crimes**," such as murder and rape.

Moreover, "young undocumented immigrants commit crime at twice the rate of young U.S. citizens." He calculates that "if undocumented immigrants committed crime nationally as they do in Arizona, in 2016 they would have been responsible for over **1,000 more murders, 5,200 rapes, 8,900 robberies, 25,300 aggravated assaults, and 26,900 burglaries**." The vast majority of the illegal immigrants Lott studies are from Mexico.

However, a number of sources (including a researcher affiliated with the Cato Institute, a non-partisan libertarian think tank) have recently challenged some of the study's methods, claiming that errors in analysis invalidate its results. Dr. Lott has responded to this critique with a defense of the study's methods, but the issue remains unresolved.

Do you think Dr. Lott's study will generate widespread media attention?

- ◯ Yes
- ◯ No

On the next screen, you will have the opportunity to authorize a $1 donation to an organization that seeks to reduce illegal immigration into the United States. We will make the donation on your behalf, so **it will not be deducted from your payment**.

>>

Donation decision

You now have the opportunity to authorize a $1 donation to **Fund the Wall**.

Fund the Wall is a nonprofit organization that seeks to reduce illegal immigration into the United States by raising money for the construction of the US-Mexico border wall. If you wish, you can choose to authorize a $1 donation to Fund the Wall. The donation **will not be deducted from your payment.**

Before proceeding, please confirm your consent to the following:

I consent to the researchers requesting the following information from my study provider: first and last name and **city. We will not have access to any other personally-identifying information, and we will use the information for study purposes only.**

- ◯ Yes, I consent and would like to proceed with the survey.
- ◯ No, I do not consent and would like to terminate the survey now.

>>

E.11.2 *No Rationale* condition

Publicizing the results of our study

As researchers, we believe it is important to communicate our findings about political and social attitudes in Bergen to the public.

If/when Dr. Lott's study is published in an academic journal, we will post the results from this survey, **including your individual donation decision and the donation decisions of all of the other respondents to this survey who consented on the previous page**, on our website.

We will promote our website via Facebook ads to Bergen residents.

What website visitors will learn

We will post your individual donation decision on our website, which will also include a description of Dr. Lott's study.

- The web page will state that **all participants were surveyed before Dr. Lott's study was published in an academic journal**
- The page lists individual donation decisions: whether or not each participant decided to authorize the donation to Fund the Wall

Donation decision

Would you like to authorize a $1 donation to **Fund the Wall**?

- ◯ Yes, I would like to authorize a $1 donation
- ◯ No, I would not like to authorize a $1 donation

Recall what people will learn when visiting the website:

- The web page will state that **all participants were surveyed before Dr. Lott's study was published in an academic journal**
- The page lists individual donation decisions: whether or not each participant decided to authorize the donation to Fund the Wall

>>

E.11.3 *Rationale* condition

Publicizing the results of our study

As researchers, we believe it is important to communicate our findings about political and social attitudes in Bergen to the public.

If/when Dr. Lott's study is published in an academic journal, we will post the results from this survey, **including your individual donation decision and the donation decisions of all of the other respondents to this survey who consented on the previous page**, on our website.

We will promote our website via Facebook ads to Bergen residents.

What website visitors will learn

We will post your individual donation decision on our website, which will also include a description of Dr. Lott's study.

- The web page will state that **all participants were shown the preliminary findings from Dr. Lott's study** before deciding whether or not to donate to Fund the Wall
- The page lists individual donation decisions: whether or not each participant decided to authorize the donation to Fund the Wall

Donation decision

Would you like to authorize a $1 donation to **Fund the Wall**?

- ◯ Yes, I would like to authorize a $1 donation
- ◯ No, I would not like to authorize a $1 donation

Recall what people will learn when visiting the website:

- The web page will state that **all participants were shown the preliminary findings from Dr. Lott's study** before deciding whether or not to donate to Fund the Wall
- The page lists individual donation decisions: whether or not each participant decided to authorize the donation to Fund the Wall

>>

E.12 Auxiliary Experiment 8: Interpretation of dissent with visible donation

E.12.1 Pre-treatment information

Do Illegal Immigrants Commit More Crime? Evidence from Dr. Lott's Study

Dr. John R. Lott, an economist formerly employed at top institutions such as Yale University and the University of Chicago, carried out a study on the relationship between illegal immigration and crime using new high-quality data. **The study has not yet been published in an academic journal, but we obtained an early version and summarize the results below.**

Dr. Lott's study uses **detailed data spanning over two decades from the Arizona state prison systems**. He finds that **"undocumented immigrants are at least 142% more likely** to be convicted of a crime than other Arizonans. They also tend to commit **more serious crimes**," such as murder and rape.

Moreover, "young undocumented immigrants commit crime at twice the rate of young U.S. citizens." He calculates that "if undocumented immigrants committed crime nationally as they do in Arizona, in 2016 they would have been responsible for over **1,000 more murders, 5,200 rapes, 8,900 robberies, 25,300 aggravated assaults, and 26,900 burglaries**. The vast majority of the illegal immigrants Lott studies are from Mexico.

However, a number of sources (including a researcher affiliated with the Cato Institute, a non-partisan think tank) have recently challenged some of the study's methods, claiming that errors in analysis invalidate its results. Dr. Lott has responded to this critique with a defense of the study's methods, but the issue remains unresolved.

>>

E.12.2 *No Rationale* condition

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

We gave your matched respondent the opportunity to authorize a $1 donation to **Fund the Wall**, a nonprofit organization that seeks to reduce illegal immigration into the United States by helping to fund and construct the US-Mexico border wall. Your matched respondent was told that their donation decision would be posted on our website. The decision on whether to authorize the donation did not have any financial consequences for your matched respondent.

Some respondents were assigned a longer version of the survey and learned about Dr. Lott's study before they decided whether or not to donate. Other respondents were assigned a shorter version of the study and **were not informed** about Dr. Lott's study before they decided whether or not to donate.

Information about your matched respondent

- Your matched respondent **was not informed about Dr. Lott's study**, which finds that illegal immigrants commit more crimes than US citizens
- Your matched respondent decided to authorize the $1 donation to Fund the Wall

>>

Why do you think your matched respondent chose to donate to Fund the Wall?

Reminder: Information about your matched respondent

- Your matched respondent **was not informed about Dr. Lott's study**, which finds that illegal immigrants commit more crimes than US citizens
- Your matched respondent decided to authorize the $1 donation to Fund the Wall

>>

After your matched respondent made their donation decision, they completed the **The Gullibility Scale**, a short questionnaire which measures **how easily people are manipulated by evidence from untrustworthy sources**.

On the next page, we will ask you to guess how your matched respondent scored on this scale. If you guess the correct option, you will be entered into a lottery for a $50 Amazon gift card.

>>

The Gullibility Scale

We administered **The Gullibility Scale**, a short questionnaire which measures **how easily people are manipulated by evidence from untrustworthy sources**, to your matched respondent.

The test is scored from 0 to 100, where 0 means "least gullible" and 100 means "most gullible". Thus, a higher score indicates that your matched respondent is more gullible.

Reminder: Information about your matched respondent

- Your matched respondent **was not informed about Dr. Lott's study**, which finds that illegal immigrants commit more crimes than US citizens
- Your matched respondent decided to authorize the $1 donation to Fund the Wall

If you had to guess, how do you think your **matched respondent** scored on **The Gullibility Scale**?

- ◯ Score between 0 and 10 (**Not at all gullible**)
- ◯ Score between 10 and 20
- ◯ Score between 20 and 30
- ◯ Score between 30 and 40
- ◯ Score between 40 and 50
- ◯ Score between 50 and 60
- ◯ Score between 60 and 70
- ◯ Score between 70 and 80
- ◯ Score between 80 and 90
- ◯ Score between 90 and 100 (**Extremely gullible**)

>>

After your matched respondent made their donation decision, they completed the **Foreign Culture Tolerance Scale**, a short questionnaire which measures **tolerance toward foreign values and traditions**.

On the next page, we will ask you to guess how your matched respondent scored on this scale. If you guess the correct option, you will be entered into a lottery for a $50 Amazon gift card.

>>

The Foreign Culture Tolerance Scale

We administered the **Foreign Culture Tolerance Scale**, a short questionnaire which measures tolerance toward **foreign values and traditions**, to your matched respondent.

The test is scored from 0 to 100, where 0 means "least tolerant" and 100 means "most tolerant". Thus, a **higher score indicates that your matched respondent is more tolerant toward foreign values and traditions.**

Reminder: Information about your matched respondent

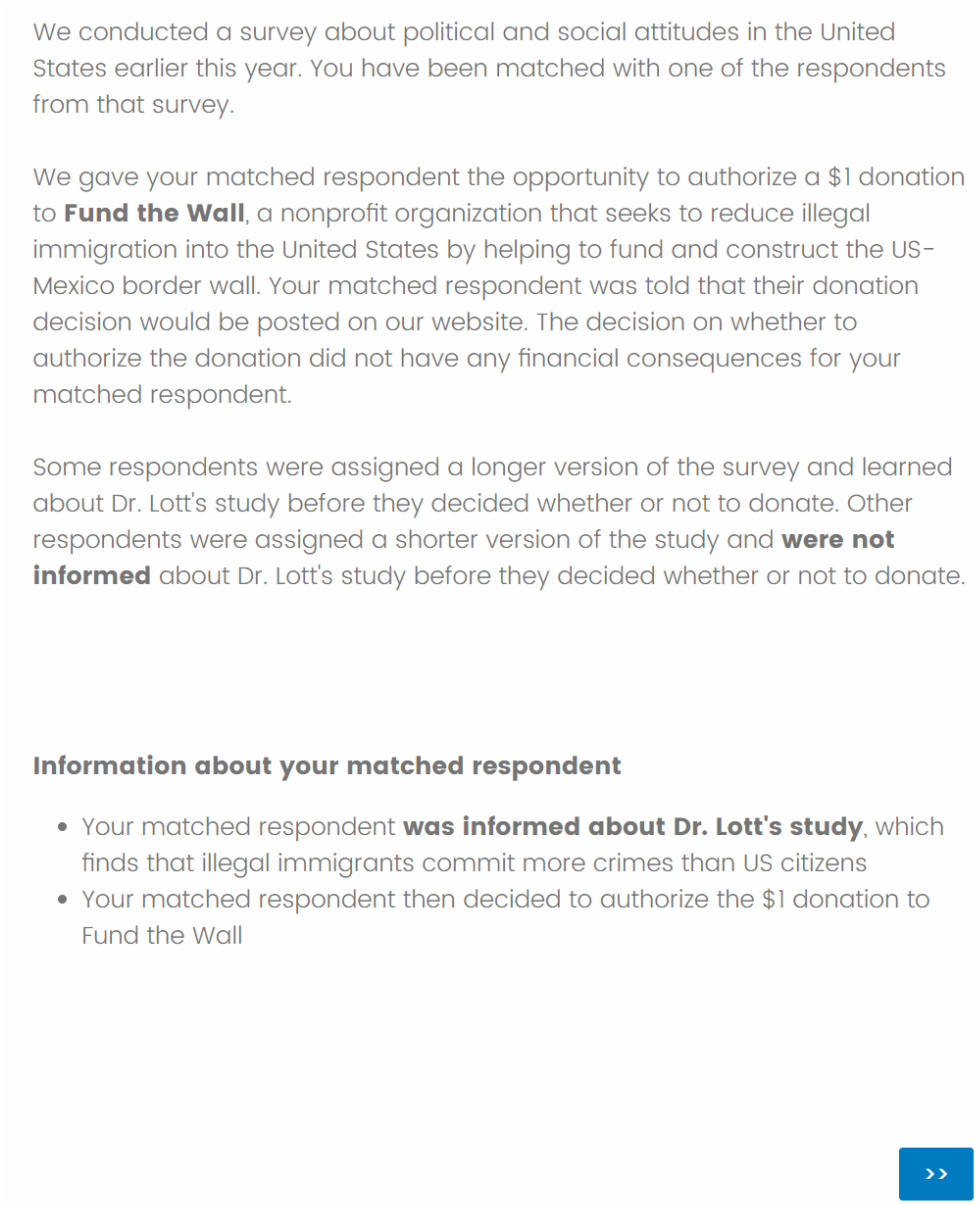
- Your matched respondent **was not informed about Dr. Lott's study**, which finds that illegal immigrants commit more crimes than US citizens
- Your matched respondent decided to authorize the $1 donation to Fund the Wall

If you had to guess, how do you think your **matched respondent** scored on the **Foreign Culture Tolerance Scale**?

- ◯ Score between 0 and 10 (**Not at all tolerant**)
- ◯ Score between 10 and 20
- ◯ Score between 20 and 30
- ◯ Score between 30 and 40
- ◯ Score between 40 and 50
- ◯ Score between 50 and 60
- ◯ Score between 60 and 70
- ◯ Score between 70 and 80
- ◯ Score between 80 and 90
- ◯ Score between 90 and 100 (**Extremely tolerant**)

>>

E.12.3 *Rationale* condition



Why do you think your matched respondent chose to donate to Fund the Wall?

Reminder: Information about your matched respondent

- Your matched respondent **was informed about Dr. Lott's study**, which finds that illegal immigrants commit more crimes than US citizens
- Your matched respondent then decided to authorize the $1 donation to Fund the Wall

>>

After your matched respondent made their donation decision, they completed the **The Gullibility Scale**, a short questionnaire which measures **how easily people are manipulated by evidence from untrustworthy sources**.

On the next page, we will ask you to guess how your matched respondent scored on this scale. If you guess the correct option, you will be entered into a lottery for a $50 Amazon gift card.

>>

The Gullibility Scale

We administered **The Gullibility Scale**, a short questionnaire which measures **how easily people are manipulated by evidence from untrustworthy sources**, to your matched respondent.

The test is scored from 0 to 100, where 0 means "least gullible" and 100 means "most gullible". Thus, a higher score indicates that your matched respondent is more gullible.

Reminder: Information about your matched respondent

- Your matched respondent **was informed about Dr. Lott's study**, which finds that illegal immigrants commit more crimes than US citizens
- Your matched respondent then decided to authorize the $1 donation to Fund the Wall

If you had to guess, how do you think your **matched respondent** scored on **The Gullibility Scale**?

- ◯ Score between 0 and 10 (**Not at all gullible**)
- ◯ Score between 10 and 20
- ◯ Score between 20 and 30
- ◯ Score between 30 and 40
- ◯ Score between 40 and 50
- ◯ Score between 50 and 60
- ◯ Score between 60 and 70
- ◯ Score between 70 and 80
- ◯ Score between 80 and 90
- ◯ Score between 90 and 100 (**Extremely gullible**)

>>

After your matched respondent made their donation decision, they completed the **Foreign Culture Tolerance Scale**, a short questionnaire which measures **tolerance toward foreign values and traditions**.

On the next page, we will ask you to guess how your matched respondent scored on this scale. If you guess the correct option, you will be entered into a lottery for a $50 Amazon gift card.

>>

The Foreign Culture Tolerance Scale

We administered the **Foreign Culture Tolerance Scale**, a short questionnaire which measures tolerance toward **foreign values and traditions**, to your matched respondent.

The test is scored from 0 to 100, where 0 means "least tolerant" and 100 means "most tolerant". Thus, a **higher score indicates that your matched respondent is more tolerant toward foreign values and traditions.**

Reminder: Information about your matched respondent

- Your matched respondent **was informed about Dr. Lott's study**, which finds that illegal immigrants commit more crimes than US citizens
- Your matched respondent then decided to authorize the $1 donation to Fund the Wall

If you had to guess, how do you think your **matched respondent** scored on the **Foreign Culture Tolerance Scale**?

- ◯ Score between 0 and 10 (**Not at all tolerant**)
- ◯ Score between 10 and 20
- ◯ Score between 20 and 30
- ◯ Score between 30 and 40
- ◯ Score between 40 and 50
- ◯ Score between 50 and 60
- ◯ Score between 60 and 70
- ◯ Score between 70 and 80
- ◯ Score between 80 and 90
- ◯ Score between 90 and 100 (**Extremely tolerant**)

>>