Reducing exclusionary attitudes through interpersonal conversation: evidence from three field experiments*

Joshua L. Kalla† David E. Broockman‡

January 17, 2020

Forthcoming, *American Political Science Review*

Abstract

Exclusionary attitudes—prejudice towards outgroups and opposition to policies that promote their well-being—are presenting challenges to democratic societies worldwide. Drawing on insights from psychology, we argue that non-judgmentally exchanging narratives in interpersonal conversations can facilitate durable reductions in exclusionary attitudes. We support this argument with evidence from three pre-registered field experiments targeting exclusionary attitudes towards unauthorized immigrants and transgender people. In these experiments, 230 canvassers conversed with 6,869 voters across 7 U.S. locations. In Experiment 1, face-to-face conversations deploying arguments alone had no effects on voters' exclusionary immigration policy or prejudicial attitudes, but otherwise identical conversations also including the non-judgmental exchange of narratives durably reduced exclusionary attitudes for at least four months ($d = 0.08$). Experiments 2 and 3, targeting transphobia, replicate these findings and support the scalability of this strategy ($ds = 0.08, 0.04$). Non-judgmentally exchanging narratives can help overcome the resistance to persuasion often encountered in discussions of these contentious topics.

*We thank the Evelyn and Walter Haas Jr. Fund, The California Wellness Foundation, The Dan and Margaret Maddox Charitable Fund, The Frist Foundation, Four Freedoms Fund, The Gateway Fund II of the Denver Foundation, The Healing Trust, The James Irvine Foundation, Luminate, and the Gill Foundation for financial support. Programmatic support was also provided by the New Conversation Initiative, Equality Federation Institute, Freedom for All Americans Education Fund, Movement Advancement Project, and the California Immigrant Policy Center. We thank seminar participants at Berkeley Haas, Columbia, the London School of Economics, the University of North Carolina, the Toronto Political Behaviour Workshop, Stanford, the University of Washington, and Yale for feedback. We also thank Rob Pressel for research assistance. All errors are our own. Replication data are available at `https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/8BFYQO`.

†Assistant Professor, Department of Political Science and Department of Statistics and Data Science, Yale University. `josh.kalla@yale.edu`, `https://joshuakalla.com`.

‡Associate Professor, Travers Department of Political Science, University of California, Berkeley. `dbroockman@berkeley.edu`, `https://polisci.berkeley.edu/people/person/david-broockman`

Exclusionary attitudes—prejudice towards outgroups and opposition to policies that promote their well-being (Enos 2014)—have been implicated in political and social strife worldwide, including populist voting in the United States (Sides, Tesler and Vavreck 2018; Reny, Collingwood and Valenzuela 2019) and the resurgence of far-right political parties in Europe (Dinas et al. 2019; Hangartner et al. 2019). Unfortunately, previous research has found that intergroup prejudices and corresponding exclusionary political attitudes typically are strong (Hopkins, Sides and Citrin 2019; Tesler 2015), arise in the presence of even minimal group differences (Tajfel 1970), persist over time (Lai et al. 2016), and are likely to further grow in response to demographic change (Velez 2018; Hajnal and Rivera 2014; Hopkins 2010; Sands and de Kadt 2019; Craig and Richeson 2014). Moreover, few strategies have been shown to allow individuals, organizations, or policymakers to feasibly reduce these exclusionary attitudes in practice (Paluck 2016; Paluck and Green 2009*b*). The few such strategies that have been identified typically decay within days (Lai et al. 2016) or require intense intervention over months or years (e.g., Paluck and Green 2009*b*; but see Broockman and Kalla 2016; Simonovits, Kezdi and Kardos 2018).

Theories from psychology suggest that individuals resist persuasion on many topics, including those related to outgroups, due to self-image concerns. These theories argue that individuals do not want to admit that their current views are in error and that yielding to persuasion may also threaten their sense of autonomy by making them feel vulnerable to manipulation by others (Cohen, Aronson and Steele 2000). Consistent with these motivations to resist persuasion, research finds that individuals engage in motivated reasoning, being motivated to dismiss evidence and arguments contrary to their views (Leeper and Slothuus 2014; Miller 1976; Sigelman and Sigelman 1984), and are often resistant to durable persuasion on many topics (e.g., Paluck 2009; Kalla and Broockman 2018).

Fortunately for individuals and organizations who wish to persuade, prior work in psychology has also documented several lab-based strategies that are able to reduce individuals' resistance to persuasion by seeking to elude or assuage these self-image concerns (e.g., Slater and Rouner 2002;

Steele, Spencer and Lynch 1993; Cohen, Aronson and Steele 2000; Chen, Minson and Tormala 2010; Itzchakov, Kluger and Castro 2017). However, it is not immediately clear how individuals and organizations seeking to persuade could practically deploy many of these lab-based strategies in the real world, such as in interpersonal conversations between colleagues or with voters as part of a political campaign.

In this paper, we argue that a strategy that attempts to address these sources of resistance to persuasion can facilitate the durable reduction of exclusionary attitudes in interpersonal conversations: the non-judgmental exchange of narratives. We define the non-judgmental exchange of narratives as *a strategy where an individual attempts to persuade another person by providing to or eliciting from them narratives about relevant personal experiences while non-judgmentally listening to the views they express*. This approach builds on two strategies in the psychology literature: narrative persuasion and high-quality listening. In this paper, we present three original, pre-registered field experiments that support our argument about the effectiveness of this approach. These experiments deployed the non-judgmental exchange of narratives to durably reduce prejudice towards two outgroups and increase support for policies that promote their well-being: unauthorized immigrants[1] and transgender people. These experiments took place across 7 U.S. locations in partnership with canvassers affiliated with 7 community-based organizations and involved conversations with 6,869 voters.

In the first field experiment we present, we randomly varied the presence of the non-judgmental exchange of narratives strategy while holding constant the other content of the conversations. This experiment found that door-to-door canvassing conversations that employed this strategy reduced exclusionary attitudes towards unauthorized immigrants for at least four months, whereas otherwise identical conversations that omitted this strategy had no detectable effects. The null effects of these otherwise identical conversations support our argument about the effects of non-judgmentally

[1] We use the term "unauthorized immigrants" because it is considered neutral and is not used by advocates on either side.

exchanging narratives, in addition to helping assuage concerns about demand effects.

Our second and third field experiments targeted attitudes towards transgender people and explored potential boundary conditions on these effects. These experiments tested whether this strategy could be effective on a new topic (in Experiments 2 and 3), with other kinds of narratives (in Experiment 2), and when these narratives are shared through other mediums (in Experiments 2 and 3). In particular, in a second treatment condition in Experiment 2, targeting transphobia, canvassers only shared narratives from a third party shown in a video; this tested whether non-judgmentally exchanging narratives from third parties could also be effective. In Experiment 3, canvassers again provided and elicited narratives to reduce transphobia, but did so by phone instead of at the door; this tested whether an in-person exchange was required. Both these experiments were motivated by a desire to test whether the non-judgmental exchange of narratives could be effective when deployed in a more easily scalable manner. Encouragingly, we found reductions in transphobia and increases in support for policies to protect transgender people from discrimination across these more scalable approaches to non-judgmentally exchanging narratives.

Our studies are relatively unique among field experiments in varying the presence of a particular strategy across multiple treatment conditions and in probing its boundary conditions in multiple experiments (e.g., Gerber, Green and Larimer 2008). Across our three experiments, we show that the strategy of non-judgmentally exchanging narratives can be successfully deployed with differing narratives; across diverse geographic contexts; when practiced by individuals and organizations with little to no prior experience; on two highly contentious topics; in the presence of contrary elite messages; and across modes of conversation. With this said, although, like many other experiments, our experiments cannot isolate a particular mechanism, we explain the theoretical reasoning that led us to expect these treatments to have the effects that they did; and, in Experiment 1, we support this reasoning by testing modified treatments where our argument predicts effects should diminish.

In the pages that follow, we first provide more theoretical background about the non-judgmental

exchange of narratives strategy we describe and detail how it was implemented in our three field experiments. We next describe the experimental design and results of our studies. We conclude by discussing broader implications and remaining questions for future research.

The Non-Judgmental Exchange of Narratives

Theoretical Background: Self-Image Concerns and Resistance to Persuasion

Theories from psychology suggest that individuals often resist persuasion because yielding to it would pose a threat to their self-image. First, yielding to persuasion may necessarily involve admitting that one has held views that were in error, threatening self-image (Cohen, Aronson and Steele 2000). Second, individuals' current attitudes may support their self-image while contrary attitudes may endanger it; for example, admitting that one's political party supports policies one opposes may threaten the self-esteem individuals derive from their partisan identities (Theodoridis 2017), as might recognizing any inconsistency between different attitudes one holds (Steele and Liu 1983; Little 2019). Such motivations may contribute to patterns well-known to political scientists, such as the pattern that individuals adopt their preferred party's positions on issues (Lenz 2013). Finally, individuals may also dislike seeing themselves as susceptible to persuasion, as this can threaten their sense of autonomy by making them feel vulnerable to manipulation by others (Brehm 1966; Slater and Rouner 2002; Pavey and Sparks 2009).

Common approaches to political persuasion that individuals and campaigns deploy may unintentionally serve to exacerbate these motives to resist persuasion. For example, campaigns often portray opponents and their supporters as deserving condemnation, as Hillary Clinton famously did in 2016 when referring to many supporters of Donald Trump's presidential candidacy as a "basket of deplorables" (Sides, Tesler and Vavreck 2018, p. 146). But such condemnations may backfire, heightening the motivation of potentially persuadable voters to counter-argue and defend

their current views. Indeed, consistent with the potential for such reactance in the case of Clinton's comment, the Trump campaign began to repeat it in campaign ads, underscoring for their supporters who might otherwise have been persuaded to vote for Clinton the threat that supporting Clinton would thus present to their self-image (Sides, Tesler and Vavreck 2018, p. 146). Likewise, in contexts such as college campuses, there is evidence for the existence of a "call-out culture" that encourages individuals to condemn perceived expressions of exclusionary attitudes (Sawaoka and Monin 2018; Lukianoff and Haidt 2019). However, while potentially playing an important role in discouraging exclusionary behavior (Paluck 2009), such condemnations may unintentionally increase resistance to persuasion among those who harbor exclusionary attitudes by heightening the negative self-image consequences of yielding to persuasion.

How can individuals and organizations seeking to persuade others attempt to overcome this challenge? It may seem obvious that condemnation would not facilitate persuasion, but it is less obvious how to reduce many sources of resistance to persuasion outside of a lab. Lab studies have highlighted a variety of strategies that reduce resistance to persuasion by reducing the threat that yielding to persuasion poses to self-image (Cohen, Aronson and Steele 2000; Sherman, Nelson and Steele 2000; Gehlbach and Vriesema 2019). For example, in some lab studies, individuals are instructed to write essays that provide alternative sources of self-esteem, such as essays reflecting on characteristics of themselves that they value (e.g., Cohen, Aronson and Steele 2000; Steele 1988). However, it is not immediately clear from these prior lab-based studies how individuals and organizations seeking to persuade others (e.g., on policies towards outgroups) could practically deploy these strategies in the real world, such as in interpersonal conversations between colleagues or with voters as part of a political campaign. It is not easy to imagine, for example, a Presidential candidate's television advertisement successfully prompting its viewers to write a reflective essay before viewing the rest of it. More generally, it is not immediately clear how individuals or organizations can argue that an opposing candidate or contrary viewpoint is incorrect without threatening the self-image of those who currently disagree with them, the very individuals they must persuade.

Strategies for Overcoming Resistance to Persuasion

The non-judgmental exchange of narratives approach we study builds on two strategies from the psychology literature for overcoming resistance to persuasion that may arise from self-image concerns.

First, previous research indicates that individuals are especially open to persuasion from narratives,[2] as prior work in psychology has found that individuals perceive narratives as less manipulative and that narratives produce less counter-arguing than direct argumentation (Green and Brock 2000; Slater and Rouner 2002; Moyer-Gusé 2008). This research finds that individuals see arguments as intended to persuade, and therefore as threatening to their sense of autonomy, but are more likely to perceive stories as primarily entertaining and non-manipulative. In addition, arguments are typically explicit (e.g., "immigrants are only a small share of the U.S. population"; e.g., Hopkins, Sides and Citrin 2019), and therefore easy for individuals to explicitly counterargue against (e.g., "but they will still compete for our jobs"). But it is more difficult to argue against a story; and individuals also often become "immersed" and "transported" into narratives, putting individuals into a less critical state of mind when they think about narratives than when individuals think about arguments, while also increasing engagement with their content (Slater and Rouner 2002; Green and Brock 2000; 2002; Moyer-Gusé 2008). Consistent with this, evidence from survey experiments finds that individuals are often more persuaded by narratives than by statistical evidence (Slater and Rouner 1996), and field experiments that successfully influence community norms through mass media often convey their messages through dramatic narratives (e.g., Paluck and Green 2009*a*; Green, Wilke and Cooper 2019; Banerjee, Barnhardt and Duflo 2017).

Second, previous research suggests that non-judgmental conversational contexts should also reduce resistance to attitude change by reducing threat to the self (Steele 1988; Cohen, Aronson and Steele 2000). Outside of lab settings, it may not be readily feasible to reduce threat to the

[2]Bilandzic and Busselle (2013) define narratives as "causally and chronologically related events played out by sentient characters."

self by prompting individuals to engage in strategies such as writing self-affirming essays. However, listening in a "non-judgmental, empathic, and respectful" manner (Itzchakov, Kluger and Castro 2017, p. 105) has been found to limit defensive reactions and increase openness to alternative viewpoints by reducing perceived threat to the self and providing affirmation (Chen, Minson and Tormala 2010; Itzchakov, Kluger and Castro 2017; see also Bruneau and Saxe 2012; Voelkel, Ren and Brandt 2019). Itzchakov, Kluger and Castro (2017) call this "high-quality listening" and we summarize it as "non-judgmental listening." In typical political exchanges where a persuader argues that one side of an issue or one candidate is superior to another, individuals' self-image may be threatened by the persuader's implicit or explicit negative judgments about individuals' existing views, and they therefore may be motivated to rebut or ignore the persuader's message. However, if a persuader shows respect by seeking out an individuals' point of view and refraining from expressing any negative judgments of it, this may affirm individuals' self-esteem and decrease the perceived threat to the self from also acknowledging the persuader's viewpoint in reciprocation (Chen, Minson and Tormala 2010; Itzchakov, Kluger and Castro 2017). In this way, creating a non-judgmental conversational context in which to persuade provides "a safe space" for political opponents to acknowledge alternative viewpoints (Itzchakov, Kluger and Castro 2017, p. 106). In addition, no viewpoint should be less threatening to the self than one's own; and so such conversations may even encourage individuals to explicitly acknowledge the merits of alternative viewpoints, promoting so-called "self-persuasion" as individuals begin to see arguments for alternative viewpoints as their own (Aronson 1999).

The non-judgmental exchange of narratives attempts to harness the strategies of narrative persuasion and non-judgmental listening identified in this prior work. Based on this prior work, we argue that interpersonal conversations that deploy the non-judgmental exchange of narratives can reduce exclusionary attitudes.

A recent paper by Broockman and Kalla (2016) lends support to this argument. Broockman and Kalla (2016) showed that conversations with 501 individuals in South Florida durably re-

duced transphobia. In these conversations, canvassers shared stories about transgender people and asked voters to share a stories about times when others judged them negatively for being different. The authors theorize that these conversations were effective because they encouraged analogic perspective-taking, a form of perspective-taking in which "perceivers try to understand the target's experience by recalling a different situation from their own experience that is presumed to parallel the target's situation" (Gehlbach and Brinkworth 2012, p. 16). However, examining the details of the canvass scripts and training from this study reveals that these conversations used several tactics that likely created a non-judgmental context and involved exchanging further narratives. Moreover, that article did not theorize—and its experiment did not manipulate—the presence of these strategies. The effects observed in Broockman and Kalla (2016) therefore could have arisen from many features of the conversations, such as the provision of basic information about who transgender people are (Flores et al. 2018). In this paper we show the presence of the non-judgmental exchange of narratives may be necessary to produce the effects they observed (in Experiment 1) and that analogic perspective-taking is itself not necessary (in Experiment 2). We also show that these same effects can be produced when non-judgmentally exchanging narratives by phone (in Experiment 3). We are not aware of other prior studies that have sought to combine narrative persuasion and non-judgmental listening.

One caveat to our argument is that it is agnostic about the content of the narratives that are exchanged, even though some narratives clearly will be more persuasive than others. In addition, different narratives may persuade through different mechanisms. In order to probe the generalizability of our argument across narratives, our empirical applications therefore show that the non-judgmental exchange of narratives can facilitate persuasion across several different kinds of narratives that likely persuade through different mechanisms. For example, Experiment 2 finds that analogic perspective-taking is not necessary to produce the effects we observe, but this may be the mechanism underpinning persuasion in Experiment 3. Likewise, none of our findings are significantly moderated by whether canvassers are members of the target outgroup, meaning that

brief contact with outgroup members is unlikely to be responsible for any of the effects we observe.[3] Further research should continue to probe boundary conditions on the effects of narratives and the mechanisms through which they can persuade.

Implementing the Non-judgmental Exchange of Narratives to Reduce Exclusionary Attitudes in Interpersonal Conversations

In this paper we test our argument that non-judgmentally exchanging narratives can facilitate durable persuasion with three experiments that focus on efforts to durably reduce exclusionary attitudes towards unauthorized immigrants and transgender people. Although future research should explore the efficacy of this strategy with other groups and issues, as we review below, attitudes towards these groups are currently highly contested in U.S. politics and thought to be strong and resistant to change.

The experiments we present study outreach from canvassers for community-based organizations who reached out to have conversations with voters in person and over the phone, common mediums of political outreach. Despite the reliable effects of high-quality personal conversations on voter turnout (Green and Gerber 2015), individuals often resist durable persuasion from these conversations (Kalla and Broockman 2018; Bailey, Hopkins and Rogers 2016), with few documented exceptions (e.g., Broockman and Kalla 2016).

In all the interpersonal conversations in our experiments, canvassers approached members of the general population by knocking on individuals' doors or calling them on the phone unannounced. Canvassers first asked individuals their view on a policy issue related to an outgroup and what considerations were on each side of the issue for them.

Next, canvassers engaged in the strategy we study: non-judgmentally exchanging narratives. To implement this strategy, canvassers provided or elicited narratives that differed across the stud-

[3]This should not be interpreted as evidence inconsistent with the "contact hypothesis," as voters' contact with canvassers met few of the conditions Allport (1954) articulated for contact that should reduce prejudice.

ies and conditions, such as narratives about personally-known outgroup members or about other personal experiences. For example, in Experiment 1, which targeted attitudes towards unauthorized immigrants, canvassers asked individuals to tell a story about "a time when someone showed [them] compassion when [they] really needed it"; per the canvass training, this was intended to help elicit "voters' own...experiences that relate to the undocumented immigrant experience." Canvassers in Experiment 1 also provided narratives about immigrants they knew or, if they were immigrants, about themselves. The canvassers' goal was to encourage individuals to engage in perspective-taking—that is, considering outgroup members' point of view (Galinsky and Moskowitz 2000; Simonovits, Kezdi and Kardos 2018)—and to activate—that is, increasing the salience of—inclusionary values (Druckman 2004*b*).

Canvassers engaged in this exchange non-judgmentally by explicitly expressing interest in understanding individuals' views and experiences, while also not expressing any negative judgments towards any statements hostile to the outgroup individuals made. The canvass training likewise instructed canvassers to "make it clear [to voters] we're not there to judge them and we're curious about their honest experience, whatever it is." During this exchange of narratives, canvassers asked questions that sought to prompt individuals to draw their own implications from the narratives. Canvassers' goal was for this non-judgmental exchange of narratives to end with individuals self-generating and explicitly stating aloud implications of the narratives that ran contrary to their previously stated exclusionary attitudes. Qualitative debriefs with the canvassers indicate that such "self-persuasion" appeared to be common.

Finally, canvassers attempted to address common misconceptions, discussed why they were supportive of inclusionary policies, and asked individuals to describe if and why the conversation changed their views.[4] The conversations lasted around 10 minutes on average. We describe more

[4]The final exercise of asking voters to rehearse any opinion change was expected to both facilitate self-persuasion, as described in the text, and also to encourage elaboration (i.e., Petty, Haugtvedt and Smith 1995). However, we did not manipulate the presence of this final rehearsal, so leave the question of whether rehearsal enhances the size and durability of the effects to future research.

details below and in the Online Appendix, where we provide the full scripts.

As mentioned above, our experiments deploy different narratives so that we can establish our findings are general across types of narratives and not driven by any one particular type of narrative. We describe the narratives exchanged in the experiments in more detail below.

Experiment 1: Does the Non-judgmental Exchange of Narratives Facilitate Reducing Exclusionary Attitudes Towards Unauthorized Immigrants?

To test whether non-judgmentally exchanging narratives facilitates durable reductions in exclusionary attitudes, we conducted a randomized field experiment targeting exclusionary attitudes towards unauthorized immigrants.

Attitudes towards unauthorized immigration are salient in contemporary American society and have important implications for immigrants' well-being (for a review, see Hainmueller and Hopkins 2014). American political elites have long used exclusionary rhetoric and supported exclusionary policies towards unauthorized immigrants, including in recent campaigns (Sides, Tesler and Vavreck 2018). The 2016 American National Election Study also found that Americans had more negative evaluations of "illegal immigrants" than of any other group asked about on the survey, including Muslims, Christian fundamentalists, and transgender people. This hostile social and political environment has undermined political support for policies that would improve unauthorized immigrants' well-being (Hainmueller et al. 2017; Hainmueller, Hangartner and Pietrantuono 2017). Prior work has found that such anti-immigrant exclusionary attitudes are strong and typically resistant to long-term change (e.g., Hopkins, Sides and Citrin 2019).

Concern about local manifestations of these trends prompted local organizations[5] to help de-

[5]These were the Tennessee Immigrant and Refugee Rights Coalition in central Tennessee; the Orange County Congregation Community Organization in Orange County, California; and Faith in the Valley in Fresno County, Cali-

velop and conduct the first intervention we report in three areas: central Tennessee; Fresno, California; and Orange County, California. In response to worksite raids by federal Immigration and Customs Enforcement (ICE) in Tennessee, a lack of legal assistance in immigration courts in Fresno, and local police reporting unauthorized immigrants to federal authorities in Orange County, the organizations had door-to-door canvassing conversations in fall 2018 in areas they expected to have higher concentrations of individuals with exclusionary attitudes towards unauthorized immigrants.[6] These groups had no prior experience attempting to reduce exclusionary attitudes through interpersonal conversations. The canvassing took place during the run-up to the 2018 US midterm elections (August – October, 2018), in which immigration issues featured prominently, such as when U.S. President Donald Trump repeatedly warned voters about a caravan of unauthorized immigrants approaching the U.S.–Mexico border.

To measure the effects of these conversations, we conducted a pre-registered, randomized, placebo-controlled experiment and parallel survey measurement using the design in Broockman, Kalla and Sekhon (2017). The experiment began by recruiting registered voters ($n = 217,600$) via mail for an ostensibly unrelated online baseline survey, presented as the first in a series of surveys not specifically about immigration and which made no reference to any potential canvassing. We gathered voters' contact information to recruit them to the survey from the public lists of registered voters, which contains a number of other covariates we use to assess the representativeness of respondents with respect to the sampling frame of registered voters we attempted to recruit. We next randomly assigned baseline survey respondents ($n = 7,870$) to Full Intervention ($n = 2,624$),

fornia.

[6]The organizations spent approximately two months preparing for the canvassing we measured, as described in more detail in the Online Appendix for Experiment 1. This preparation included an approximately six week period of qualitative "iteration" on the script. During this period, canvassers attempted different conversational approaches and narrative prompts with voters not in the study and debriefed their experiences with the candidate prompts in regular conference calls with the group leaders, a team from the New Conversation Initiative, and the researchers. For example, one candidate prompt was to ask voters about a time when they showed someone else compassion; canvassers felt this did not generate as much understanding of the experience of unauthorized immigrants as the prompt ultimately selected. This period also allowed canvassers to be trained in the skills of non-judgmental listening and eliciting narratives, as well as the experimental procedures.

Abbreviated Intervention ($n = 2,623$), or Placebo conditions ($n = 2,623$). Blocked random assignment was conducted at the household level ($n = 6,551$ households), such that participants within the same household were always assigned to the same experimental condition.

Next, to deliver the intervention, staff and volunteers affiliated with the partner organizations went door-to-door during August – October, 2018 to visit individuals' homes at their addresses in the voter registration database. As described above, canvassers began by knocking on voters' doors unannounced. Canvassers then asked to speak with the person on their list who had enrolled in the study and confirmed the person's identity. After the person's identity was confirmed, canvassers implemented the experimental condition corresponding with the person's random assignment.

When individuals were assigned to the Full Intervention, the conversations proceeded as described in the introduction: canvassers asked individuals for their view on the issue, engaged in the non-judgmental exchange of narratives, addressed common misconceptions, and made supportive arguments. The Full Intervention condition included the non-judgmental exchange of narratives on two topics: canvassers' and individuals' previous experience with immigrants and, second, as described above, about "a time when someone showed [them] compassion when [they] really needed it." Canvassers were trained to particularly focus on the latter. These narratives were intended to promote general perspective-taking (Galinsky and Moskowitz 2000), analogic perspective-taking (Gehlbach and Brinkworth 2012), and the salience of compassion as a value (Rokeach 1971).

The Abbreviated Intervention condition removed the exchange of these narratives but was otherwise identical to the Full Intervention, including containing addressing common misconceptions and making supportive arguments, similar to a traditional political canvass.

The Placebo condition was a brief (approximately 1 minute) conversation unrelated to immigration, conducted solely for the purpose of identifying which individuals could be contacted (Nickerson 2005).[7]

The Online Appendix provides further details about the intervention, including the full scripts.

[7]These were news consumption in Tennessee, gun violence in Fresno, and housing in Orange County.

Table 1: Summary of differences between conditions and results in previous study and Experiments 1-3

Study	Broockman and Kalla (2016)	Experiment 1		Experiment 2		Experiment 3
Topic	Transphobia	Unauthorized Immigrants		Transphobia		Transphobia
Condition Name	Full Intervention	Full Intervention	Abbreviated Intervention	Participants' and Video Narratives	Video Narratives Only	Participants' Narratives by Phone
Intervention Contents						
Non-judgmental exchange of narratives...						
○ From participants (voter and canvasser)	YES	YES	NO	YES	NO	YES
○ In video	YES	NO	NO	YES	YES	NO
Address concerns and deliver talking points	YES	YES	YES	YES	YES	YES
Results						
ITT†	Positive effects ($d = 0.16$, $p < 0.001$)	Positive effects ($d = 0.08$, $p < 0.001$)	Null effects ($d = 0.02$, $p = 0.27$), statistically distinguishable from Full Intervention ($d = 0.06$, $p < 0.01$)	Positive effects ($d = 0.08$, $p < 0.001$)	Positive effects ($d = 0.08$, $p < 0.001$)	Positive effects ($d = 0.04$, $p < 0.001$)
CACE‡	$d = 0.22$	$d = 0.12$	$d = 0.03$ (Abbreviated vs. Placebo)	$d = 0.10$	$d = 0.10$	$d = 0.08$

Notes: Each Experiment also contained a Placebo condition not shown in the Table. These placebo conditions contained no persuasive content on the topics but are used as a baseline for comparison when estimating the effect sizes shown in the Table.

†To summarize the results of each study, we first average the pre-specified Overall Index in each study across survey waves to compute a pooled Overall Index. We then report intent-to-treat (ITT) effects on this pooled Overall Index, which represents the mean difference between individuals assigned to each condition among all individuals who identified themselves at their doors, regardless of whether the conversation continued after that point. The ITT estimates represent the average causal effect of attempting to treat people who open their doors, even if they refuse to converse soon after. This means the ITT estimates are "diluted" by the presence of individuals who open the door but do not enter into the conversation.

‡To estimate the implied Complier Average Causal effect (CACE), or the effect among those who received the intervention, we estimate compliance under a conservative definition of compliance, whether participants got to the "first rating" part of the conversation where they initially told canvassers how they felt about the policy. The CACE estimates represent the average causal effect of treating the people who do enter (or would have entered) into the conversation. These estimates require the assumption that there was no effect of beginning the conversation but not reaching this "first rating." The p-values are identical to the ITT results.

Table 1 also summarizes the experimental conditions.

Canvassers successfully reached 2,374 individuals at their doors across the three conditions. Approximately 70% of voters assigned to the Full Intervention condition who were reached went on to complete the entire conversation and 77% shared a personal narrative with the canvasser, as recorded by canvassers after each conversation ended. On average, voters who identified themselves at their doors in the Full Intervention condition went on to converse for 11 minutes on average; this figure is 5 minutes for voters in the Abbreviated Intervention condition.[8] The canvassers had no experience conducting in-person conversations to reduce exclusionary attitudes prior to the project, had an average age of 25, and were ethnically diverse, with 54% self-identifying as Latino.

We recruited individuals who were reached to follow-up surveys that began 4 days ($n = 1,578$), 30 days ($n = 1,508$), and 3-6 months ($n = 1,384$) after the conversations.[9] We monitored responses to an open-ended question about any comments on the survey and debriefed the canvassers to see if participants registered any suspicions that the canvass intervention was related to the surveys and found none.

The Online Appendix include further recruitment, design, survey, and estimation details, representativeness assessments (Table OA1), and tests of design assumptions such as the proper implementation of the placebo, balance checks, and checks for differential attrition (Tables OA2-8). The Online Appendix reports that endline participants are slightly more likely to be older, white, and to politically participate than individuals in the sampling frame recruited to the baseline sur-

[8]We expected the Abbreviated Intervention to be shorter, as its name suggests, because this condition removed the non-judgmental exchange of narrative strategy. This may raise the question of whether the increased duration of the interaction confounds our interpretation of the results. However, we do not find that a longer interaction is more effective in Experiment 2. In addition, any alternative comparison condition that held duration constant while removing the non-judgmental exchange of narratives would necessarily need to introduce some other additional content, leading to a different confound. For example, if we had included additional arguments in the Abbreviated Intervention, we would not be able to tell whether the Full Intervention was more effective because the particular arguments used were less effective than the particular stories used in the Full Intervention. We thus followed the approach in Gerber, Green and Larimer (2008) (in which some of the treatments were also longer than others) of removing particular components of the treatment we theoretically expected to increase its effects without replacing them with alternatives.

[9]The first two survey waves were done on a rolling basis after each canvass took place. The final survey was launched on the same day for all participants, regardless of the date they were canvassed. For the average participant, the final survey wave was completed approximately 4.5 months after they were canvassed (sd of 0.5 months).

vey, patterns that also appear in Experiments 2 and 3. These patterns appear to bias the estimates downwards, as Table OA25 in the Online Appendix shows that applying survey weights typically increases the point estimates.

The intervention sought to reduce exclusionary attitudes towards unauthorized immigrants along two pre-registered dimensions: increasing support for more inclusionary government policies (e.g., granting legal status to people who were brought to the US illegally as children) and decreasing prejudice towards unauthorized immigrants, defined broadly as negative attitudes towards the group (e.g., "I would have no problem living in areas where undocumented immigrants live"). The surveys included 6 items measuring support for policies related to immigrants and 7 items capturing anti-immigrant prejudice. As we pre-registered, we combine these two groups of items into two indices, a policy index as a prejudice index, as well as a third overall index containing all 13 items.

To estimate treatment effects in all our experiments, we use linear regressions including pre-registered pre-treatment covariates to increase precision (Gerber and Green 2012). Given the household-level random assignment in all our studies, the standard errors are clustered at the household level. We pre-registered this estimation strategy and the covariates we would use to increase precision.

These estimated treatment effects are intent-to-treat (ITT) effects among all individuals who open their doors and identify themselves before the intervention and placebo scripts diverge. Because not all individuals continue with the intervention after this point, the estimates are therefore diluted by the presence of individuals who did not receive the entire intervention (Gerber and Green 2012). As the Online Appendix describes, and as shown in Table 1, complier average causal effect (CACE) estimates that correct for this by estimating the effects among those who do enter the conversation are larger.

Experiment 1 Results

Figure 1 shows the results.

The first panel shows that the Full Intervention increased support for inclusive policies as measured in the surveys 1 week ($d = 0.11$, $t = 4.12$, $p < 0.001$), 1 month ($d = 0.06$, $t = 2.39$, $p < 0.02$), and 3-6 months after the intervention ($d = 0.08$, $t = 2.78$, $p < 0.01$). Averaging individuals' responses at all three points in time, the pooled effect is also significant ($d = 0.09$, $t = 3.89$, $p < 0.001$).[10] Examining results on dichotomized versions of the individual items in the policy index,[11] the average share of inclusive policies individuals strongly supported increased from 29% in the Placebo condition to 33% in the Full Intervention condition ($p < 0.01$). For example, while the Abbreviated Intervention had no effect on individuals strongly supporting granting legal status to people who were brought to the US illegally as children and who have graduated from a U.S. high school, individuals assigned to the Full Intervention were 4.7 percentage points more likely to indicate strong support ($p < 0.04$). Likewise, when dichotomizing the policy items by whether individuals supported each policy at all, instead of expressing indifference or opposition, the share of policies individuals supported at all increased by 2.2 percentage points in the Full Intervention condition ($p = 0.058$). Note again that all these estimates are intent-to-treat estimates and that the compliance-adjusted estimates would be larger. See Online Appendix Tables OA23-4 for additional results on the dichotomized individual policy items.

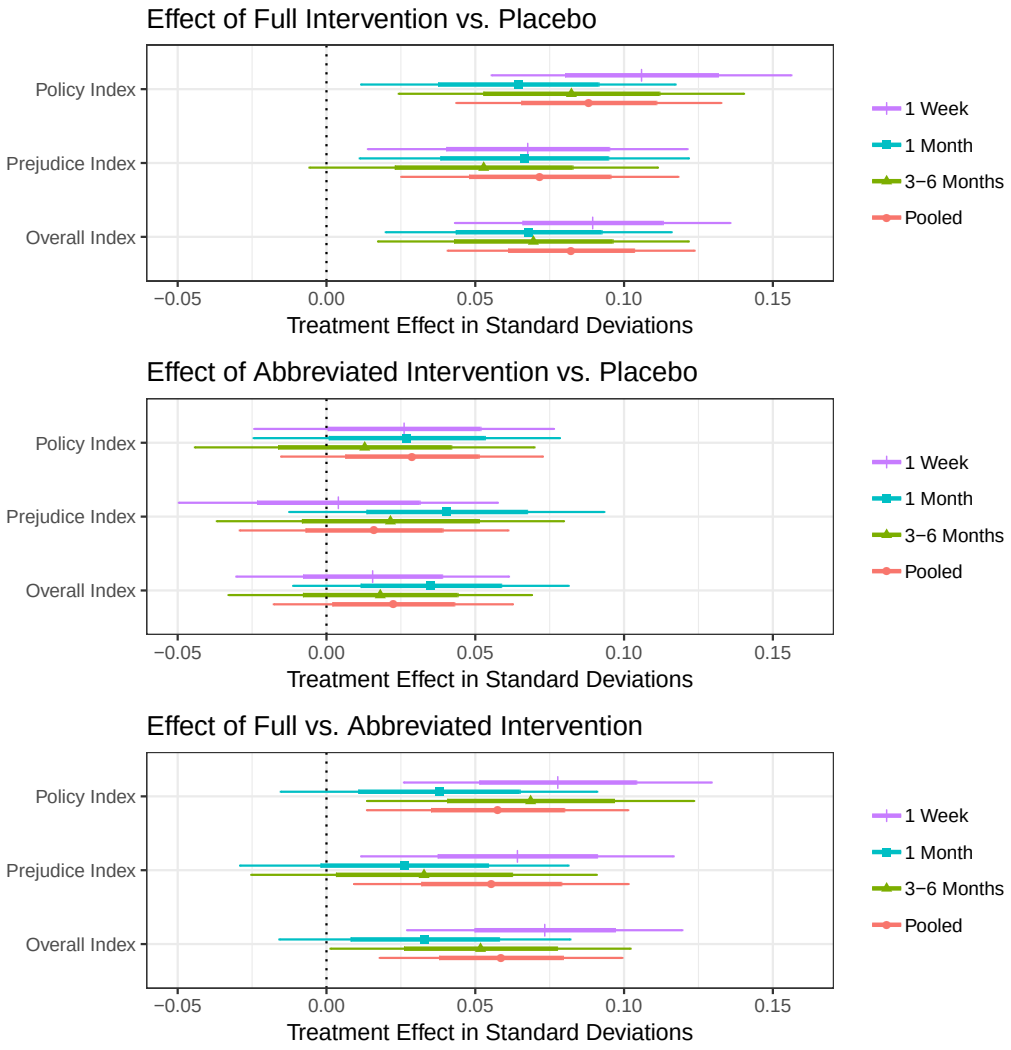
The Full Intervention also reduced prejudice towards unauthorized immigrants in the surveys 1 week ($d = 0.07$, $t = 2.47$, $p < 0.02$), 1 month ($d = 0.07$, $t = 2.36$, $p < 0.02$), and 3-6 months after the intervention ($d = 0.05$, $t = 1.77$, $p < 0.08$; pooled estimate $d = 0.07$, $t = 3.02$, $p < 0.01$).

The second panel of Figure 1 shows that the Abbreviated Intervention, which excluded the non-

[10] We did not pre-register how to summarize the results across multiple survey waves, but choose to compute a simple average of individuals' responses to multiple survey waves to limit our discretion.

[11] These analyses of dichotomized versions of the individual items were not pre-registered; we conducted them to help illustrate the substantive size of the effects. We exclude the compassion item from these analyses of the dichotomized items because it is not a specific policy akin to a ballot measure or candidate policy position. The effects are largest on this item, so including it would strengthen the results. See discussion surrounding Tables OA23-4 in the Online Appendix.

Figure 1: Experiment 1 Results: Intent-to-Treat Effects



Notes: Each panel shows the estimated intent-to-treat effects when comparing the two experimental conditions described in the panel title (e.g., the top panel compares the Full Intervention condition to the Placebo condition). Within each panel, we show treatment effects on the pre-specified primary outcome indices. Results are average treatment effects with 1 standard error (thick) and 95% confidence intervals (thin). To form each pooled index, we average each respondent's values for the corresponding index across all post-treatment survey waves. See Online Appendix Tables OA9-11 for numerical point estimates and standard errors.

judgmental exchange of narratives, had effects indistinguishable from zero. This is consistent with the positive results of the Full Intervention not being driven by demand effects; the experimental design is capable of producing null results.

However, as the third panel in Figure 1 shows, we can statistically distinguish the effects of the Full from the Abbreviated intervention, the most direct test of the impact of the non-judgmental exchange of narratives. This indicates that including the non-judgmental exchange of narratives significantly increased the treatment effects. Those assigned to the Full instead of Abbreviated Intervention were significantly more supportive of inclusive policies in the surveys 1 week ($d = 0.08$, $t = 2.95$, $p < 0.01$), 1 month ($d = 0.04$, $t = 1.40$, $p = 0.17$), and 3-6 months after the intervention ($d = 0.07$, $t = 2.45$, $p < 0.02$). The pooled result is $d = 0.06$ ($t = 2.57$, $p < 0.01$). There are largely similar results for the prejudice index in the surveys 1 week ($d = 0.06$, $t = 2.40$, $p < 0.02$), 1 month ($d = 0.03$, $t = 0.93$, $p = 0.36$), and 3-6 months after the intervention ($d = 0.03$, $t = 1.11$, $p = 0.27$); averaging individuals' responses at these three points in time, the average effect on the prejudice index is statistically significant ($d = 0.05$, $t = 2.36$, $p < 0.02$).

Online Appendix Tables OA9-11 present the precise point estimates, standard errors, t-statistics, and p-values. Note that all these statistics use our pre-specified estimation approach of incorporating pre-treatment covariates to increase precision, which is central to the experimental design we employed (Broockman, Kalla and Sekhon 2017). Online Appendix Tables OA9-11 also present results without covariates for transparency; as one would expect, without incorporating covariates, the standard errors are larger, as are the p-values.

There was little meaningful treatment effect heterogeneity by canvasser or voter attributes; the conversations were broadly persuasive regardless of which canvassers or voters were involved. Online Appendix Table OA12 shows that the effects of the Full Intervention are similar regardless of whether the canvasser is an immigrant ($d = 0.12$, $t = 2.20$, $p < 0.03$) or is not an immigrant ($d = 0.08$, $t = 3.12$, $p < 0.01$). The clearly significant effects for non-immigrant canvassers mean the effects cannot be attributed to mere contact and that voters need not be prompted to take

canvassers' own perspective for the intervention to be effective. Table OA16 also shows the Full Intervention was effective when implemented by both Latino and non-Latino canvassers. Tables OA17-25 present additional heterogeneous treatment effect results, including by voter education, economic well-being, race, and partisanship. There are few clear patterns of heterogeneity, although there are clearly significant persuasive effects among both Republican and Independent voters. In the Online Appendix we show that there is no evidence of differential attrition by condition (Tables OA7 and OA8) and that applying survey weights if anything increases the point estimates (Table OA25).

To summarize, Experiment 1 has three important findings. First, interpersonal conversations that deployed the non-judgmental exchange of narratives reduced exclusionary attitudes towards unauthorized immigrants—a widely discussed, openly stigmatized group, attitudes towards whom have been deemed strong and resistant to change (Hopkins, Sides and Citrin 2019). Second, these effects lasted for at least 4.5 months in a competitive political context (the immediate run-up to the 2018 U.S. midterm elections) in which elites, including U.S. President Donald Trump, expressed contrary policy arguments and open hostility towards the group; and these effects persisted even among self-identified Republicans. Third, we experimentally demonstrated that the non-judgmental exchange of narratives was primarily responsible for generating these effects, as removing it significantly reduces if not eliminates the effects of these conversations.

Experiment 2: Probing Boundary Conditions With a Door-to-door Canvass Targeting Transphobia

Experiment 2 targets exclusionary attitudes towards transgender people. As with policies towards unauthorized immigrants, policies towards transgender people have been increasingly salient in recent years, with U.S. President Donald Trump issuing a Memorandum preventing transgender people from serving in the military and legislators in sixteen states introducing laws in 2017 requir-

ing transgender people to use the bathroom of the sex they were assigned at birth (Kralik 2017).

Experiment 2 attempts to replicate our findings, explore potential boundary conditions, and assess a more scalable version of the non-judgmental exchange of narratives strategy. In particular, Experiment 2 includes a Video Narratives Only condition where canvassers showed voters a video narrative about a third party but did not supply their own narratives nor elicit individuals' narratives. To share this video narrative, canvassers showed and discussed a video displayed on canvasser's smartphones that depicts a transgender woman unknown to the canvassers and the participants describing a time when a restaurant manager attempted to force her to use the men's restroom but other patrons intervened, allowing her to use the restroom of her choosing.[12]

The Video Narratives Only condition in Experiment 2 allows us to test the generality of Experiment 1's findings by testing whether non-judgmentally exchanging narratives can be effective when different narratives are used which do not include the canvasser sharing their own narrative nor eliciting a narrative from the voter. Recall that the Full Intervention in Experiment 1 involved eliciting narratives from voters and canvassers sharing narratives about their own experiences with voters; but the Video Narratives Only condition in Experiment 2 does neither. This condition therefore allows us to test whether it is necessary to elicit narratives from voters or for canvassers to share their own narratives in order for narratives to persuade. Second, recall that in the Full Intervention in Experiment 1, canvassers also all shared narratives about immigrants they personally knew or, if the canvassers were immigrants, about themselves. Experiment 2's Video Narratives Only condition probes whether hearing a narrative about an outgroup member from that outgroup member or someone who personally knows them may be required to produce these effects. Experiment 2's Video Narratives Only condition does so by omitting narratives about outgroup members from conversation participants and only including the video narrative about a third party unknown to either the canvasser or participant.

In addition to the Video Narratives Only condition, Experiment 2 also included a Participants'

[12]The video is publicly available at `https://www.youtube.com/watch?v=YNwVrWGQneg`.

and Video Narratives condition. The video narratives described above were also present in this condition. However, when individuals were assigned to the Participants' and Video Narratives condition, canvassers also shared their own narratives and elicited narratives from voters about experiences with outgroup members and personal experiences of being treated differently, narratives we expected to further promote the salience of inclusionary values, perspective-taking, and analogic perspective-taking in particular. This condition allows us to benchmark the effects of the Video Narratives Only condition against a condition similar to the Full Intervention in Experiment 1. Table 1 again summarizes the conditions.

To measure the effects of these interventions, we again conducted a pre-registered randomized placebo-controlled experiment and parallel survey measurement. The experiment took place in 2016 in four areas: Atlanta, Georgia; Cleveland, Ohio; Jacksonville, Florida; and Scottsdale, Arizona. First, we recruited registered voters ($n = 324,620$) via mail for an ostensibly unrelated online baseline survey, presented as the first in a series of surveys. These surveys were broad university-sponsored surveys that included dozens of items unrelated to transphobia to disguise their connection with the upcoming intervention. We next randomly assigned respondents to this baseline survey ($n = 8,456$) to either the Participants' and Video Narratives condition ($n = 2,815$), the Video Narratives Only condition ($n = 2,817$), or a Placebo condition receiving a brief conversation about banning plastic bags, an issue unrelated to transphobia ($n = 2,824$). Blocked random assignment was conducted at the household level ($n = 3,485$ households), such that participants within the same household were always assigned to the same experimental condition.

Next, canvassers affiliated with four partner non-profit organizations[13] visited individuals' homes at their addresses in the voter registration database. When study participants were assigned to the Participants' and Video Narratives condition, the intervention proceeded similarly to as de-

[13]These were Equality Foundation of Georgia in Atlanta, Georgia; Equality Ohio Education Fund in Cleveland, Ohio; Equality Florida Institute in Jacksonville, Florida; and ONE Community in Scottsdale, Arizona.

scribed above. The scripts are available in the Online Appendix. As described above, in the Video Narratives Only condition, canvassers continued discussing the narratives non-judgmentally, but did not provide their own narratives or ask for voters' narratives, instead only showing and discussing the narrative of the third party in a video. Consistent with the canvassers successfully implementing this change, in conversations that successfully began, records the canvassers made after each conversation indicate that voters and canvassers ultimately shared their own stories 69% and 85% of the time, respectively, when voters were assigned to the Full Intervention. These figures are only 12% and 19% when voters were assigned to the Video Narratives Only condition. On average, individuals who identified themselves at the door in the Participant and Video Narratives condition conversed for 10.5 minutes on average; this figure is 7.7 minutes in the Video Narratives Only condition. 37% of conversations were conducted by canvassers who identify as transgender.

Canvassers successfully reached 1,858 individuals at their doors across the three conditions. We recruited individuals who were reached to follow-up surveys that began one week ($n = 1,044$) and one month ($n = 989$) after the conversations. We monitored survey responses and debriefed canvassers to see if participants had any suspicions that the canvass intervention was related to the surveys and found none.

The intervention sought to reduce transphobia along two pre-registered dimensions: increasing support for more inclusionary government policies (e.g., support for "a law in your state that would protect gay and transgender people from discrimination in employment, housing, and public accommodations") and decreasing prejudice towards transgender people (e.g., "I would support a friend choosing to have a sex change"). Each survey included 9 items measuring support for policies related to transgender people and 6 items capturing anti-transgender prejudice. As we pre-registered, we combine these two groups of items into two indices, a policy index and a prejudice index, as well as a third index containing all 15 items.

The Online Appendix includes further recruitment, design, survey, and estimation details, tests of design assumptions (such as the proper implementation of the placebo, balance checks, and

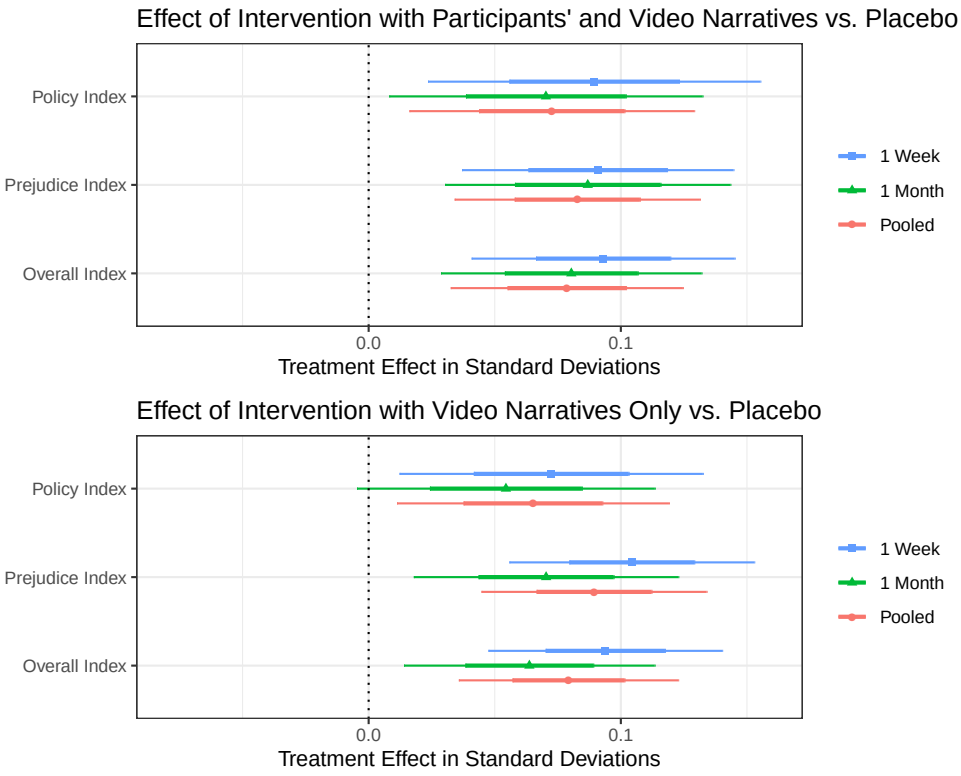
checks for differential attrition; see Online Appendix Tables OA28-31, OA36-7), representativeness assessments (Table OA26), and estimates with survey weights, which are typically slightly larger (Table OA50).

Experiment 2 Results

Figure 2 shows the results. The first panel shows that the Participants' and Video Narratives condition successfully increased support for inclusive policies as measured in the surveys 1 week ($d = 0.09$, $t = 2.66$, $p < 0.01$) and 1 month ($d = 0.07$, $t = 2.22$, $p < 0.03$) after the intervention (pooled effect $d = 0.07$, $t = 2.52$, $p < 0.02$). It also reduced prejudice towards transgender people in the surveys 1 week ($d = 0.09$, $t = 3.31$, $p < 0.001$) and 1 month ($d = 0.09$, $t = 3.01$, $p < 0.001$) after the intervention (pooled effect $d = 0.08$, $t = 3.34$, $p < 0.001$).

However, the Video Narratives condition that involved the non-judgmental exchange of narratives shown in videos but did not include participants' own narratives was also effective. In particular, the Video Narratives Only intervention also successfully increased support for inclusive policies as measured in the surveys 1 week ($d = 0.07$, $t = 2.36$, $p < 0.02$) and 1 month ($d = 0.05$, $t = 1.81$, $p < 0.07$; pooled effect $d = 0.07$, $t = 2.37$, $p < 0.02$). The Video Narratives Only intervention also reduced prejudice towards transgender people in the surveys 1 week ($d = 0.10$, $t = 4.21$, $p < 0.001$) and 1 month ($d = 0.07$, $t = 2.63$, $p < 0.01$) after the intervention (pooled effect $d = 0.09$, $t = 3.93$, $p < 0.001$). Online Appendix Tables OA40-42 present the precise point estimates, standard errors, t-statistics, and p-values. (All differences between the two treatment conditions in Experiment 2 were insignificant. Online Appendix Tables OA40-42 report the point estimates and standard errors on this difference; although we can be confident that both treatment conditions had effects, the standard error on the differences in their effects is large, meaning we also cannot rule out the possibility of meaningful differences between the conditions.) Note that all these statistics use our pre-specified estimation approach of incorporating pre-treatment covariates to increase precision. Online Appendix Tables OA40-42 also present results without covariates;

Figure 2: Experiment 2 Results: Intent-to-Treat Effects



Notes: Each panel shows the estimated treatment effects when comparing the two experimental conditions described in the panel title (e.g., the top panel compares the Participants' and Video Narratives condition to the Placebo condition). Within each panel, we show treatment effects on the pre-specified primary outcome indices. Results are average treatment effects with 1 standard error (thick) and 95% confidence intervals (thin). To form each pooled index, we average each respondent's values for the corresponding index across all post-treatment survey waves. See Online Appendix Tables OA40-42 for numerical point estimates and standard errors.

without incorporating covariates, the standard errors are larger, as are the p-values.

One sign that new attitudes are strong is that they endure over time; another is that they resist attack (Petty, Haugtvedt and Smith 1995). In Experiment 1, we found durable persuasive effects despite the presence of contrary elite messages from U.S. President Donald Trump during the 2018 midterm elections. In Experiment 2, lacking such a naturally-occurring context, we provided contrary messages in our survey. In particular, we showed an opposing advertisement mid-way through the post-treatment surveys and pre-registered that we would separately analyze indices of

items asked before and after the opposing video was shown. (This video was shown to all participants in both the treatment and control groups.) Consistent with these new attitudes formed from the canvassing treatment being strong, we find that the treatment effects are essentially identical on the index of items asked after individuals were shown the opposition advertisement (see Online Appendix Tables OA43-4). This is also propitious for the external validity for our results to a competitive political context.

In Table OA52 we show that the canvassing treatments had effects regardless of whether delivered by transgender or cisgender canvassers; and in Table OA56 we show consistent results across participants' partisan identifications. Additional subgroup analyses are presented in Tables OA53-55. Tables OA59-60 also show results on the dichotomized policy items, which are broadly consistent with both creating new supporters and strengthening support.

However, as shown in Table 1, we also note that the two interventions reported here were around one-half as effective as that reported in Broockman and Kalla (2016). We pre-registered an expectation that this was a less favorable implementation context than in Broockman and Kalla (2016) given that the partner organizations had less prior experience implementing longer canvassing interactions, which could explain this smaller treatment effect.

In summary, Experiments 1 and 2 find that non-judgmentally exchanging narratives present in a video (Experiment 2) and narratives from participants in the conversation (Experiment 1, where no video was present) are both able to durably reduce exclusionary attitudes.

Experiment 3: Probing Scalability With Phone Conversations Targeting Transphobia

Our third field experiment administered a version of the intervention in which individuals non-judgmentally exchanged narratives over the phone. Canvassers could not show voters videos over the phone; therefore, similar to the Full Intervention condition in Experiment 1, the intervention

only included canvasser- and voter-supplied narratives (again, see Table 1 for summary). The prompts used to elicit canvasser- and voter-supplied narratives in Experiment 3 were the same as those used in Experiment 2 (narratives about experiences with outgroup members and about personal experiences of being treated differently). Experiment 3 therefore both further replicates the finding from Experiment 1 that a conversation including only participants' narratives (and no video narratives) can have durable effects and shows they generalize to the less personal, more easily scalable context of a telephone conversation. These narratives were intended to promote the salience of inclusionary values, perspective-taking, and especially analogic perspective-taking, which Experiment 2 found was not necessary for persuasion but could nevertheless still have persuasive effects.

This experiment took place in the same four areas as Experiment 2, among individuals who either lived outside of the canvass area or whose household members were never reached during the canvass phase (e.g., no one was home when a canvasser knocked). We randomized these participants to a treatment group targeted with the Participants' Narratives by Phone condition ($n = 6,879$) or a Placebo condition receiving a brief telephone call unrelated to transphobia ($n = 6,888$). Random assignment was conducted at the household level ($n = 12,081$ households), such that participants within the same household always received the same experimental condition. Next, canvassers called individuals on the phone and administered either the Placebo or Participants' Narratives by Phone condition. Canvassers successfully reached 2,637 individuals. Individuals in the Participants' Narratives by Phone condition who were reached on the phone conversed for 6.6 minutes on average. We recruited individuals who were reached to follow-up surveys that began one week ($n = 1,943$) and one month ($n = 1,897$) after the conversations. These follow-up surveys asked the same questions as in Experiment 2, and we formed the same policy and prejudice indices in the same manner.

The Online Appendix includes further recruitment, design, survey, and estimation details, tests of design assumptions (such as the proper implementation of the placebo, balance checks, and

checks for differential attrition; see Tables OA32-35, 37, 39), representativeness assessments (Table OA27), and estimates with survey weights, which are similarly sized (Table OA51).

Experiment 3 Results

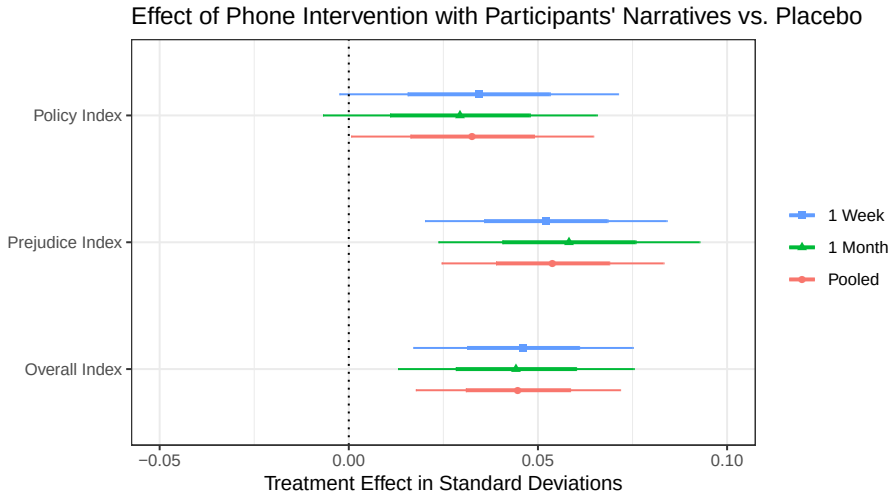
Figure 3 shows the results. The Participants' Narratives by Phone intervention reduced prejudice towards transgender people in the surveys 1 week ($d = 0.05$, $t = 3.20$, $p < 0.001$) and 1 month ($d = 0.06$, $t = 3.31$, $p < 0.001$) after the intervention (pooled effect $d = 0.05$, $t = 3.60$, $p < 0.001$). The intervention also likely increased support for inclusive policies after the intervention; although the effects measured in the 1 week ($d = 0.03$, $t = 1.83$, $p < 0.07$) and 1 month ($d = 0.03$, $t = 1.59$, $p = 0.11$) surveys do themselves not reach statistical significance, the pooled effect on policy attitudes averaging the two surveys does ($d = 0.03$, $t = 2.00$, $p < 0.05$). Online Appendix Tables OA45-8 present the precise point estimates, standard errors, t-statistics, and p-values. Note that all these statistics use our pre-specified estimation approach of incorporating pre-treatment covariates to increase precision. Online Appendix Tables OA45-8 also present results without covariates for transparency; without incorporating covariates, the standard errors are larger, as are the p-values.

As in Experiment 2, we also again see that the new attitudes the intervention formed are resistant to attack, as the results are similar on an index of items asked after individuals were shown an opposition advertisement (see Online Appendix Tables OA48-9). In Tables OA57-8 we show that the intervention was broadly effective across participants' partisan identifications and levels of political knowledge. Tables OA61-2 also show results on the dichotomized policy items.

Discussion

Prejudice towards outgroups and opposition to policies that promote their well-being have contributed to social and political challenges worldwide. Individuals and organizations that wish to

Figure 3: Experiment 3 Results: Intent-to-Treat Effects



Notes: This Figure shows the estimated treatment effects when comparing the Phone Intervention with Participants' Narratives condition to the Placebo condition. We show treatment effects on the pre-specified primary outcome indices. Results are average treatment effects with 1 standard error (thick) and 95% confidence intervals (thin). To form each pooled index, we average each respondent's values for the corresponding index across all post-treatment survey waves. See Online Appendix Tables OA45-47 for numerical point estimates and standard errors.

reduce these exclusionary attitudes, be they individuals speaking with acquaintances or political campaigns seeking to change voter opinion, have few proven strategies available to them to productively engage those who disagree with them on these topics. If they do engage, social norms may also encourage individuals to engage in strategies such as condemnation and argumentation that may in fact be counterproductive (Itzchakov, Kluger and Castro 2017) and lead individuals to believe others do not respect them (Cramer 2016; 2012). Meanwhile, existing strategies largely have effects that rapidly decay or require sustained intervention over months or years (Paluck and Green 2009*b*; Lai et al. 2016).

Our results—which focus on two highly stigmatized groups and divisive political issues—indicate that individuals and organizations can durably reduce exclusionary attitudes in these interpersonal conversations by non-judgmentally exchanging narratives. Our evidence shows that this strategy can be effective across varied contexts: these interventions were successfully deployed to

complete strangers in the general population across seven sites by seven different organizations; we found effects when administering this strategy on an extremely salient issue in the midst of many contrary elite messages (for Experiment 1, immigration during the 2018 US midterm elections; and in Experiments 2 and 3, from an opposing advertisement shown in the survey); we found them regardless of whether narratives were shared through the mediums of in-person conversation (Experiments 1 and 2), phone conversation (Experiment 3), or video (Experiment 2); and from narratives of different types, including when participants exchanged personal narratives (Experiments 1, 2, and 3) and when canvassers shared a narrative from a third party (Experiment 2). Our findings therefore suggest optimism that individuals seeking to reduce exclusionary attitudes may be able to productively employ this strategy in everyday interpersonal conversations.

The contexts in which these experiments took place also suggest optimism for efforts for individuals and organizations to implement the non-judgmental exchange of narratives at scale: none of the seven organizations we worked with had previously implemented such an intervention, nor had the canvassers had any such prior experience. Previous research has found smaller treatment effects of other interventions when they are implemented at larger scale and by new partner organizations (Allcott 2015; Grossman, Humphreys and Sacramone-Lutz 2019), consistent with the smaller effects we found in this study than in Broockman and Kalla (2016), as shown in Table 1.[14] While future research should continue to test potential boundary conditions on these effects, our findings already suggest optimism for other practitioners seeking to implement our findings. The fact that the canvassers themselves had no prior experience also underscores the normative benefits of deliberations between citizens (Druckman and Nelson 2003; Druckman 2004*a*; Landemore

[14]For example, Broockman and Kalla (2016) collaborated with an organization in South Florida with extensive experience in such canvassing, raising questions about external validity to organizations with less experience. However, all the experiments in this paper were conducted in collaboration with groups with no prior experience with canvassing to reduce exclusionary attitudes. Accordingly, our pre-registration for Experiments 2 and 3 indicated that we viewed this as "a much less favorable implementation context" than the South Florida context. As noted above, we expect this relative inexperience is responsible for the smaller treatment effects seen in Experiments 2 and 3 than in Broockman and Kalla (2016), as shown in Table 1. Experiment 1 also targeted immigration attitudes, which may be more crystallized and difficult to change than attitudes towards transgender people.

2013) and suggests that Americans may be able to adopt this strategy in their deliberations with others.

At the same time, we do not wish to overstate the substantive size of the effects we estimated. On the one hand, some may see these effects as relatively sizable given the null effects of many other door-to-door persuasion programs (Kalla and Broockman 2018) and the difficulty of changing attitudes, at least on immigration, in many survey-based experiments (Hopkins, Sides and Citrin 2019, although see effects even larger than those observed here in Simonovits, Kezdi and Kardos (2018)). On the other hand, many social psychologists would traditionally consider effect sizes of the sizes we observed (intent-to-treat effects of $d = 0.08$ in Experiments 1 and 2 and $d = 0.04$ in Experiment 3) small. Moreover, given the size of the effects we observe, a campaign implementing this approach should expect that a very large number of such conversations would be needed to produce detectable changes in aggregate public opinion or changes in electoral outcomes. At the same time, a campaign looking for strategies to change aggregate public opinion may have no choice but to pursue strategies with small effects; few if any other campaign tactics have been rigorously shown to have lasting meaningful effects in the field on public opinion.

Another important limitation of this work, as with many experiments, is that we were unable to test all the specific mechanisms that might produce the reduction in exclusionary attitudes that we observe. For example, it is difficult to control what processes individuals engage in (e.g., perspective-taking, activation of inclusionary values, or other emotional processes) when supplying their own narratives outside a laboratory setting. Although we detailed our theoretical reasoning and were able to support this reasoning by testing modified treatments where our argument predicts effects should diminish, further tests of these mechanisms could be taken up by future studies.

However, our findings nevertheless are notable for pinpointing a strategy that is important to generate the effects we observed. Most importantly, in Experiment 1, removing the non-judgmental exchange of narratives significantly reduced if not eliminated the effectiveness of the intervention, supporting our argument that this strategy facilitates the reduction of exclusionary attitudes. We

also conceptually replicated these findings using different forms of narratives in the context of conversations that took place through different modes in Experiments 2 and 3. The results of all three experiments also are inconsistent with the alternative explanations that the physical presence of outgroups are necessary or sufficient for the effects we observed.

With this said, future work should continue to refine these interventions and our understanding of why they work. Five areas seem especially important. First, future work should explore how to apply the non-judgmental exchange of narratives in mass media (Paluck 2009; 2010), where limiting defensive reactions through non-judgmental listening may prove more difficult but narratives may still be effective. Second, it is an open question whether this strategy would be effective when targeting attitudes on other topics where personal narratives may be more difficult to share and elicit (e.g., climate change). Third, our theoretical argument is agnostic to the type of narratives shared. Although we showed that the effects we observed are not particular to any one type of narrative, future research should seek to better understand which narrative strategies are most effective for different types of issues, voters, and contexts; no doubt some narratives would fail to persuade on some topics (e.g., as occurred in an experiment on door-to-door canvassing on abortion, reported in Broockman, Kalla and Sekhon 2017, Section 6). Fourth, what consequences would result if both sides of an issue engaged in this strategy, especially in a traditional partisan campaign? Although any competing efforts to change policy attitudes may cancel out, such efforts may still increase tolerance for those who share opposing viewpoints (Mutz 2002; Bruneau and Saxe 2012). Finally, it would also be valuable to test what if any behavioral consequences such conversations have, such as on actual voting behavior or prejudiced behaviors (Sands 2017; Enos 2016), as well as any potential effects on implicit, as opposed to explicit, attitudes (Lai et al. 2016).

Our results also suggest a possible tension between strategies for reducing exclusionary attitudes at the individual level and strategies for reducing their behavioral consequences at a societal level. Previous field experiments find that promulgating norms that discourage exclusionary behaviors—i.e., signaling that exclusionary behaviors will be judged negatively by others—can

effectively reduce the consequences of intergroup prejudice, even though this does not reduce exclusionary attitudes themselves (Paluck 2009). However, our work joins others in suggesting that signaling individuals will *not* be judged negatively for expressing exclusionary attitudes may facilitate their openness to changing these attitudes (Itzchakov, Kluger and Castro 2017). Efforts to promote a culture where individuals expect social opprobrium for engaging in exclusionary behavior may therefore need to balance the value of creating conditions in which individuals do not feel threatened by discussing their attitudes and experiences with those who wish to persuade them.

References

Allcott, Hunt. 2015. "Site selection bias in program evaluation." *The Quarterly Journal of Economics* 130(3):1117–1165.

Allport, Gordon W. 1954. *The Nature of Prejudice*. Cambridge, MA: Addison-Wesley.

Aronson, Elliot. 1999. "The power of self-persuasion." *American Psychologist* 54(11):875–884.

Bailey, Michael A., Daniel J. Hopkins and Todd Rogers. 2016. "Unresponsive and unpersuaded: The unintended consequences of a voter persuasion effort." *Political Behavior* 38(3):713–746.

Banerjee, Abhijit, Sharon Barnhardt and Esther Duflo. 2017. Movies, Margins, and Marketing: Encouraging the Adoption of Iron-Fortified Salt. In *Insights in the Economics of Aging*, ed. David A. Wise. University of Chicago Press pp. 285–306.

Bilandzic, Helena and Rick Busselle. 2013. Narrative persuasion. In *The Sage handbook of persuasion: Developments in theory and practice*, ed. Lijiang Shen and James Price Dillard. Thousand Oaks, CA: Sage pp. 200–219.

Brehm, Jack W. 1966. *A theory of psychological reactance.* Oxford: Academic Press.

Broockman, David E. and Joshua L. Kalla. 2016. "Durably reducing transphobia: A field experiment on door-to-door canvassing." *Science* 352(6282):220–224.

Broockman, David E., Joshua L. Kalla and Jasjeet S. Sekhon. 2017. "The design of field experiments with survey outcomes: A framework for selecting more efficient, robust, and ethical designs." *Political Analysis* 25(4):435–464.

Bruneau, Emile G. and Rebecca Saxe. 2012. "The power of being heard: The benefits of 'perspective-giving'in the context of intergroup conflict." *Journal of experimental social psychology* 48(4):855–866.

Chen, Frances S., Julia A. Minson and Zakary L. Tormala. 2010. "Tell me more: The effects of expressed interest on receptiveness during dialog." *Journal of Experimental Social Psychology* 46(5):850–853.

Cohen, Geoffrey L., Joshua Aronson and Claude M. Steele. 2000. "When beliefs yield to evidence: Reducing biased evaluation by affirming the self." *Personality and Social Psychology Bulletin* 26(9):1151–1164.

Craig, Maureen A. and Jennifer A. Richeson. 2014. "More diverse yet less tolerant? How the increasingly diverse racial landscape affects white Americans' racial attitudes." *Personality and Social Psychology Bulletin* 40(6):750–761.

Cramer, Katherine J. 2012. "Putting inequality in its place: Rural consciousness and the power of perspective." *American Political Science Review* 106(3):517–532.

Cramer, Katherine J. 2016. *The politics of resentment: Rural consciousness in Wisconsin and the rise of Scott Walker*. University of Chicago Press.

Dinas, Elias, Konstantinos Matakos, Dimitrios Xefteris and Dominik Hangartner. 2019. "Waking

Up the Golden Dawn: Does Exposure to the Refugee Crisis Increase Support for Extreme-Right Parties?" *Political Analysis* 27(2):244–254.

Druckman, James N. 2004*a*. "Political preference formation: Competition, deliberation, and the (ir) relevance of framing effects." *American Political Science Review* 98(4):671–686.

Druckman, James N. 2004*b*. "Priming the vote: Campaign effects in a US Senate election." *Political Psychology* 25(4):577–594.

Druckman, James N. and Kjersten R. Nelson. 2003. "Framing and deliberation: How citizens' conversations limit elite influence." *American Journal of Political Science* 47(4):729–745.

Enos, Ryan D. 2014. "Causal effect of intergroup contact on exclusionary attitudes." *Proceedings of the National Academy of Sciences* 111(10):3699–3704.

Enos, Ryan D. 2016. "What the demolition of public housing teaches us about the impact of racial threat on political behavior." *American Journal of Political Science* 60(1):123–142.

Flores, Andrew R. et al. 2018. "Challenged expectations: Mere exposure effects on attitudes about transgender people and rights." *Political Psychology* 39(1):197–216.

Galinsky, Adam D. and Gordon B. Moskowitz. 2000. "Perspective-taking: decreasing stereotype expression, stereotype accessibility, and in-group favoritism." *Journal of personality and social psychology* 78(4):708–724.

Gehlbach, Hunter and Christine Calderon Vriesema. 2019. "Meta-bias: A practical theory of motivated thinking." *Educational Psychology Review* 31(1):65–85.

Gehlbach, Hunter and Maureen Elizabeth Brinkworth. 2012. "The social perspective taking process: Strategies and sources of evidence in taking another's perspective." *Teachers College Record* 114:1–29.

Gerber, Alan S. and Donald P. Green. 2012. *Field experiments: Design, analysis, and interpretation*. W. W. Norton.

Gerber, Alan S., Donald P. Green and Christopher W. Larimer. 2008. "Social pressure and voter turnout: Evidence from a large-scale field experiment." *American political Science review* 102(1):33–48.

Green, Donald P. and Alan S. Gerber. 2015. *Get out the vote: How to increase voter turnout*. Brookings Institution Press.

Green, Donald P., Anna Wilke and Jasper Cooper. 2019. "Countering violence against women at scale: A mass media experiment in rural Uganda." Working paper, available at `https://www.poverty-action.org/sites/default/files/publications/GreenWilkeCooper2019.pdf`.

Green, Melanie C. and Timothy C. Brock. 2000. "The role of transportation in the persuasiveness of public narratives." *Journal of personality and social psychology* 79(5):701–721.

Green, Melanie C. and Timothy C. Brock. 2002. In the mind's eye: Transportation-imagery model of narrative persuasion. In *Narrative impact: Social and cognitive foundations*, ed. Melane C. Green, Jeffrey J. Strange and Timothy C. Brock. Mahwah, NJ: Lawrence Erlbaum Associates Publishers pp. 315–341.

Grossman, Guy, Macartan Humphreys and Gabriella Sacramone-Lutz. 2019. "Information technology and political engagement: Mixed evidence from Uganda." *Journal of Politics* .

Hainmueller, Jens and Daniel J. Hopkins. 2014. "Public attitudes toward immigration." *Annual Review of Political Science* 17:225–249.

Hainmueller, Jens, Dominik Hangartner and Giuseppe Pietrantuono. 2017. "Catalyst or crown:

Does naturalization promote the long-term social integration of immigrants?" *American Political Science Review* 111(2):256–276.

Hainmueller, Jens et al. 2017. "Protecting unauthorized immigrant mothers improves their children's mental health." *Science* 357(6355):1041–1044.

Hajnal, Zoltan and Michael U. Rivera. 2014. "Immigration, Latinos, and white partisan politics: The new democratic defection." *American Journal of Political Science* 58(4):773–789.

Hangartner, Dominik, Elias Dinas, Moritz Marbach, Konstantinos Matakos and Dimitrios Xefteris. 2019. "Does Exposure to the Refugee Crisis Make Natives More Hostile?" *American Political Science Review* 113(2):442–455.

Hopkins, Daniel J. 2010. "Politicized places: Explaining where and when immigrants provoke local opposition." *American political science review* 104(1):40–60.

Hopkins, Daniel J., John Sides and Jack Citrin. 2019. "The muted consequences of correct information about immigration." *Journal of Politics* 81(1):315–320.

Itzchakov, Guy, Avraham N. Kluger and Dotan R. Castro. 2017. "I am aware of my inconsistencies but can tolerate them: The effect of high quality listening on speakers' attitude ambivalence." *Personality and Social Psychology Bulletin* 43(1):105–120.

Kalla, Joshua L. and David E. Broockman. 2018. "The minimal persuasive effects of campaign contact in general elections: Evidence from 49 field experiments." *American Political Science Review* 112(1):148–166.

Kralik, Joellen. 2017. "Bathroom Bill Legislative Tracking." National Conference of State Legislatures.

Lai, Calvin K. et al. 2016. "Reducing implicit racial preferences: II. Intervention effectiveness across time." *Journal of Experimental Psychology: General* 145(8):1001–16.

Landemore, Hélène. 2013. "On minimal deliberation, partisan activism, and teaching people how to disagree." *Critical Review* 25(2):210–225.

Leeper, Thomas J. and Rune Slothuus. 2014. "Political parties, motivated reasoning, and public opinion formation." *Political Psychology* 35:129–156.

Lenz, Gabriel S. 2013. *Follow the leader?: how voters respond to politicians' policies and performance*. University of Chicago Press.

Little, Andrew T. 2019. "The distortion of related beliefs." *American Journal of Political Science* 63(3):675–689.

Lukianoff, Greg and Jonathan Haidt. 2019. *The coddling of the American mind: How good intentions and bad ideas are setting up a generation for failure*. Penguin Books.

Miller, Richard L. 1976. "Mere exposure, psychological reactance and attitude change." *Public Opinion Quarterly* 40(2):229–233.

Moyer-Gusé, Emily. 2008. "Toward a theory of entertainment persuasion: Explaining the persuasive effects of entertainment-education messages." *Communication Theory* 18(3):407–425.

Mutz, Diana C. 2002. "Cross-cutting social networks: Testing democratic theory in practice." *American Political Science Review* 96(1):111–126.

Nickerson, David W. 2005. "Scalable protocols offer efficient design for field experiments." *Political Analysis* 13(3):233–252.

Paluck, Elizabeth Levy. 2009. "Reducing intergroup prejudice and conflict using the media: a field experiment in Rwanda." *Journal of personality and social psychology* 96(3):574.

Paluck, Elizabeth Levy. 2010. "Is it better not to talk? Group polarization, extended contact, and perspective taking in eastern Democratic Republic of Congo." *Personality and Social Psychology Bulletin* 36(9):1170–1185.

Paluck, Elizabeth Levy. 2016. "How to overcome prejudice." *Science* 352(6282):147.

Paluck, Elizabeth Levy and Donald P. Green. 2009*a*. "Deference, dissent, and dispute resolution: An experimental intervention using mass media to change norms and behavior in Rwanda." *American political science review* 103(4):622–644.

Paluck, Elizabeth Levy and Donald P. Green. 2009*b*. "Prejudice reduction: What works? A review and assessment of research and practice." *Annual review of psychology* 60:339–367.

Pavey, Louisa and Paul Sparks. 2009. "Reactance, autonomy and paths to persuasion: Examining perceptions of threats to freedom and informational value." *Motivation and Emotion* 33(3):277–290.

Petty, Richard E., Curtis P. Haugtvedt and Stephen M. Smith. 1995. Elaboration as a determinant of attitude strength: Creating attitudes that are persistent, resistant, and predictive of behavior. In *Attitude strength: Antecedents and consequences*, ed. R.E. Petty and J.A. Krosnick. Lawrence Erlbaum Associates, Inc. pp. 93–130.

Reny, Tyler T., Loren Collingwood and Ali Valenzuela. 2019. "Vote Switching in the 2016 Election: How Racial and Immigration Attitudes, Not Economics, Explain Shifts in White Voting." *Public Opinion Quarterly* 83(1):91–113.

Rokeach, Milton. 1971. "Long-range experimental modification of values, attitudes, and behavior." *American psychologist* 26(5):453.

Sands, Melissa and Daniel de Kadt. 2019. "Segregation drives racial voting: New evidence from South Africa." *Political Behavior* .

Sands, Melissa L. 2017. "Exposure to inequality affects support for redistribution." *Proceedings of the National Academy of Sciences* 114(4):663–668.

Sawaoka, Takuya and Benoît Monin. 2018. "The paradox of viral outrage." *Psychological science* 29(10):1665–1678.

Sherman, David K., Leif D. Nelson and Claude M. Steele. 2000. "Do messages about health risks threaten the self? Increasing the acceptance of threatening health messages via self-affirmation." *Personality and Social Psychology Bulletin* 26(9):1046–1058.

Sides, John, Michael Tesler and Lynn Vavreck. 2018. *Identity crisis: The 2016 presidential campaign and the battle for the meaning of America*. Princeton University Press.

Sigelman, Lee and Carol K. Sigelman. 1984. "Judgments of the Carter-Reagan debate: The eyes of the beholders." *Public Opinion Quarterly* 48(3):624–628.

Simonovits, Gábor, Gabor Kezdi and Peter Kardos. 2018. "Seeing the World Through the Other's Eye: An Online Intervention Reducing Ethnic Prejudice." *American Political Science Review* 112(1):186–193.

Slater, Michael D. and Donna Rouner. 1996. "Value-affirmative and value-protective processing of alcohol education messages that include statistical evidence or anecdotes." *Communication Research* 23(2):210–235.

Slater, Michael D. and Donna Rouner. 2002. "Entertainment—education and elaboration likelihood: Understanding the processing of narrative persuasion." *Communication Theory* 12(2):173–191.

Steele, Claude M. 1988. The psychology of self-affirmation: Sustaining the integrity of the self. In *Advances in experimental social psychology*. Vol. 21 Elsevier pp. 261–302.

Steele, Claude M., Steven J. Spencer and Michael Lynch. 1993. "Self-image resilience and dissonance: The role of affirmational resources." *Journal of personality and social psychology* 64(6):885–96.

Steele, Claude M. and Thomas J. Liu. 1983. "Dissonance processes as self-affirmation." *Journal of personality and social psychology* 45(1):5–19.

Tajfel, Henri. 1970. "Experiments in intergroup discrimination." *Scientific American* 223(5):96–103.

Tesler, Michael. 2015. "Priming predispositions and changing policy positions: An account of when mass opinion is primed or changed." *American Journal of Political Science* 59(4):806–824.

Theodoridis, Alexander G. 2017. "Me, myself, and (I),(D), or (R)? Partisanship and political cognition through the lens of implicit identity." *Journal of Politics* 79(4):1253–1267.

Velez, Yamil Ricardo. 2018. "Residential Mobility Constraints and Immigration Restrictionism." *Political Behavior* pp. 1–25.

Voelkel, Jan G, Dongning Ren and Mark Brandt. 2019. "Political inclusion reduces political prejudice." Working paper, available at `https://psyarxiv.com/dxwpu/`.

Online Appendix for Experiment 1 (Unauthorized Immigration)

Joshua Kalla[*]
David Broockman[†]

Contents

[*]Yale University, Departments of Political Science and Statistics & Data Science, josh.kalla@yale.edu
[†]University of California, Berkeley, Department of Political Science, dbroockman@berkeley.edu

The full replication code and data that produces this report will be available at https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/8BFYQO.

This experiment was pre-registered at https://egap.org/registration/5138.

Intervention Details

Training

Before beginning the experiment, the partner organizations dedicated 278 unique canvasser shifts to developing the intervention and developing the training. The canvassers in this experiment were paid. They received training when they first started, including following a more experienced canvasser for a day. Throughout the program, they received ongoing training and feedback. The trainings focused on providing canvassers with the skills to listen to and ask questions of voters in a non-judgmental manner that would elicit narratives from voters about their experiences. Trainings often involved role play and viewing video of past canvass conversations. Trainings were led by the New Conversation Initiative.

Canvasser Demographics

The canvassers for this project were primarily paid canvassers recruited by the three local organizations. Neither the canvassers nor the local organizations had prior experience conducting in-person conversations to reduce exclusionary attitudes. 77% of the conversations were conducted by canvassers age 30 and under (average age was 25), 60% by female canvassers, 54% by canvassers who self-identified as Latino, and 24% by canvassers who self-identified as immigrants.

Intervention Procedure

The canvassers were trained to follow the below procedure when approaching homes when subjects were in the treatment conditions. Being mainly concerned with external validity, this procedure does not strictly rely on only one theoretical paradigm as is common in lab studies. However, the majority of the time in the training and in the conversations was spent on how to non-judgmentally exchange narratives. (The full scripts are reproduced below.)

Canvassers themselves were not aware of the details of the experiment or the survey and nowhere in the conversation did they indicate that the effects of the conversation were being measured or part of the study.

Establish Contact

1. **Determine if voter is home.** The canvasser knocks on the door and says, "Hi, I'm [canvasser's name]. Are you [subject's name]?" If the subject identifies themself, the canvasser marks "Voter came to door" on their walk list. This leads the voter to be targeted for resurveying. Note that this first step is identical in the placebo and treatment conditions.

Create Non-Judgmental Context

2. **Intervention begins: inform subject about the policy being discussed.** The canvassers began the intervention by engaging in a series of strategies to elicit participants' opinions in a non-judgmental manner. First, canvassers informed voters that they were at the door to discuss a policy related to unauthorized immigrants (e.g., in Tennessee, "Based on what you know now, would you say you are against, undecided, or in favor of large-scale arrests and detainment of undocumented immigrants at their place of work?"). Canvassers then asked voters about their opinion on the policy and then asked them to explain their position. Canvassers were trained to ask these questions in a non-judgmental

manner, not indicating they were pleased or displeased with any particular answer, but rather to appear genuinely interested in hearing the subject ruminate on the question. This was intended to encourage effortful reflection and to build rapport.

Exchange Narratives - Removed in the abbreviated condition

3. **Exchange narratives about personal experience with immigration.** The canvasser then asked the voter if they know anyone who is an immigrant and, in particular, an unauthorized immigrant. If the voter knows someone, the canvasser would have the voter talk about how they know this person, their immigration story, and how it must feel to be an immigrant. Whether or not the voter knows an immigrant, the canvasser would always share their immigration story. This might be a personal story or about a friend or family member. The canvasser would end this section by asking the voter if there is anything about the story that they can relate to, encouraging perspective taking [5].
4. **Exchange narratives about a personal experience with compassion.** The intervention attempted to prompt values that would lead participants to be more supportive of unauthorized immigrants and encourage analogic perspective-taking. To do this, canvassers asked voters to share a time when someone showed them compassion. If necessary, canvassers sometimes told their own stories of being shown compassion in order to make voters feel comfortable sharing a story of their own. For many canvassers, this would involve telling a story about being shown compassion that related to their own experiences as immigrants or as close friends or family to immigrants.

Canvassers' goal was for this non-judgmental exchange of narratives to end with individuals self-generating and explicitly stating aloud implications of the narratives that ran contrary to their previously stated exclusionary attitudes.

Address Concerns

5. **Address voter concerns**. At this point, the canvasser would return to any concerns about unauthorized immigrants that the voter may have mentioned earlier. The canvasser would talk through these concerns and, where applicable, provide talking points to refute them. Canvassers were trained not to address concerns until this point in the conversation so that voters would not feel threatened by this section. Only after rapport had been established, stories shared, and the value of compassion activated would canvassers address concerns. Canvassers would address concerns at this point in both the Full and Abbreviated interventions.

Make the Case

6. **Provide arguments and information**. The canvasser would then reiterate for the voter why they were canvassing and why they hoped the voter would become more supportive of unauthorized immigrants.

Encourage Active Processing

7. **Ask for opinion again; rehearse opinion change.** The intervention ended with canvassers asking voters if and why the conversation changed their exclusionary attitudes towards unauthorized immigrants. Rehearsal of opinion change is a strategy that has been shown to facilitate active processing and increase the persistence of attitude changes [13]. The canvasser then thanked the subject and left.

Placebo Procedure

The sole purpose of the placebo conversations was to identify voters who were home and thus voters with whom the intervention could be plausibly attempted (see section entitled "Placebo Design"). When approaching homes where subjects were in the placebo group, canvassers followed the following procedure instead:

1. **Determine if voter is home.** The canvasser knocks on the door and says, "Hi, I'm [canvasser's name]. Are you [subject's name]?" If the subject identifies themself, the canvasser marks "Voter came to door" on their walk list. This leads the voter to be targeted for resurveying. Note that this first step is identical in the placebo and treatment conditions.
2. **Placebo begins.** The canvasser would then deliver the placebo survey, which varied by site.
3. **Conversation ends.** The canvasser thanks the subject and leaves.

Scripts

Below are the scripts for the Full Intervention, Abbreviated Intervention, and Placebo conditions.

Survey Recruitment Procedures and Experimental Design

In this section we describe the survey recruitment procedures and the experimental design. We assess the representativeness of the sample at each step and test design assumptions in other sections.

Baseline Survey

To measure the effects of the intervention, we conducted ostensibly unrelated surveys of voters living in two regions of California and one region of Tennessee. To recruit voters to these surveys, the partner organizations first provided us with contact information for voters living in the areas they planned to canvass, acquired from the publicly available list of registered voters. We invited these voters to the baseline survey by mail. The survey was called the "UNIVERSITY-UNIVERSITY Public Opinion Study". (See more detail in "Additional Survey Details" section.)

The recruitment letter included the survey web URL, a unique login for each voter, and instructions for taking the survey online. To participate, respondents entered the URL from the letter in their computer or smartphone and then their unique login. We mailed letters to 122712 households that contained individual logins for 217600 people; when multiple eligible voters lived in the same household, we sent the household one letter that contained a unique login for each person.

Voters were offered no incentives for completing the baseline survey but were offered $5 for completing each follow-up survey. Voters received these incentives via email (collected during the survey) immediately upon completion of the survey. Voters could redeem these $5 incentives as gift-cards to Amazon, iTunes, Starbucks, WalMart, or Home Depot or as donations to Habitat for Humanity, the National Parks Foundation, or Clean Water Fund.

Random Assignment of Households

7870 voters completed the baseline survey and provided a valid email address. We randomly assigned one-third to each treatment condition. Voters were randomly assigned at the household level, ensuring that multiple voters who completed the pre-survey within the same household were always assigned to the same treatment condition. All analyses adjust standard errors to account for this clustered assignment (see details below). This procedure is identical to that used in [2] and follows the best practices for field experiments with survey outcomes [3].

The household-level clustered random assignment took place within blocks of three households. These blocks were formed by matching households on household size and on household-level-average values of baseline covariates (a factor measuring baseline views on immigration and a factor measuring baseline views on partisan politics). Within each block, one household (cluster) was assigned to each condition. This pre-treatment blocking reduces the chance of imbalance between conditions and improves precision [3].

Random Assignment of Turfs

On the day of each canvass, groups of households were formed into "turfs" by the staff at the partner organizations. "Turfs" are groups of nearby households convenient for two canvassers to visit by walking a short distance. Households were put in groups blind to treatment assignment and simply based on the geographic layout of households to be canvassed that day. A route connecting the households in the turf were then drawn, again blind to treatment assignment, such that an efficient route could be followed; half of the households were marked for a Canvasser A and half for a Canvasser B in an order each canvasser could follow. The groups of households (turf) were then randomly assigned to pairs of canvassers by having canvassers pick a number corresponding to a turf out of a hat. Then, canvass leaders flipped a coin to determine which canvasser would knock on A doors and which on B doors. In some cases, Canvasser A and B would be one person. Data-quality checks conducted after the canvass ensured that canvassers all properly

canvassed the assigned doors within their turf. This random assignment of canvassers to turf allows us to assess canvasser-level treatment effect heterogeneity, such as by canvasser immigration status.

Placebo Design for Delivering Intervention

Canvassers attempted to have a conversation about unrelated issues with voters in the placebo group and a conversation containing the intervention with voters in the two treatment groups. In Fresno, the placebo was to get voters to sign a petition on gun violence; in Orange County the placebo was a brief survey on rent control; in Tennessee, the placebo was a brief survey on media consumption. This placebo-controlled experimental design [12] is common in studies of door-to-door canvassing interventions and field experiments more generally [3]. Nickerson [12] summarizes the placebo design:

> Rather than rely upon a control group that receives no attempted treatment, the group receiving the placebo can serve as the baseline for comparison for the treatment group... assuming that (1) the two treatments have identical compliance profiles; (2) the placebo does not affect the dependent variable; and (3) the same type of person drops out of the experiment for the two groups.

Gerber et al. [7] similarly summarize the design:

> subjects who agree to participate in a study and for whom the prospect of treatment is imminent are randomly assigned to receive either the treatment or the placebo.

The sole purpose of the placebo discussion was to identify subjects who were home and thus with whom a conversation at the door could be attempted (versus subjects who were not home at all or would not even open the door). Identifying this group allows a direct comparison of subjects with whom the intervention actually began to subjects with whom the intervention could have begun but did not because of their random assignment (and thus with whom a conversation about recycling began instead). This design dramatically improves the precision of door-to-door canvassing experiments [3].

We implemented the placebo design as follows.

First, the canvassers began by implementing an identical procedure regardless of experimental condition. Canvassers were given walk lists of voters to contact that had been sequentially ordered by voters' addresses blind to voters' treatment assignment. Canvassers proceeded down the list of houses in the experiment in this order, knocking on one door after another without regard to the household's experimental group. The beginning of the conversation was also identical in each condition: "Hi, are you [subject's name]?" If the subject identified him/herself or came to the door at this point, the canvasser then checked a box called "Voter came to door" on their walk list. The experimental sample consists of those who came to the door at this point.

Only after canvassers determined whether the voter they were looking for came to the door or not did they begin either implementing the intervention or delivering the placebo. Importantly, nothing was different in the procedure before this point: voters did not know the canvasser intended to have a conversation with them about immigration or the placebo issues before identifying themselves or not; canvassers did not inform voters about the topic of the conversation before this point.

These procedures guarantee an unbiased experimental comparison among voters who came to the door and then were delivered the intervention or were then not delivered the intervention based on their random assignment [12, 3].

One strength of this study's research design is that we are able to sensitively test the placebo design's key assumption: the kinds of voters who identify themselves at their doors before the placebo starts and before the intervention starts are similar. Our tests support this assumption. We describe these tests in the *Tests of Design Assumptions* section.

Follow-Up Surveys

Following the placebo design described above, we conducted multiple waves of follow-up surveys for voters who came to the door in any condition. These follow-up surveys began around 1 week, 1 month, and 3-6 months after the day each voter was canvassed. We solicited voters to complete these re-surveys at the email addresses they provided in the baseline survey. Three reminders to complete the follow-up surveys were sent for each survey wave.

Note that to the extent any voters answered the wrong surveys or did not answer the surveys carefully, this measurement error would lead us to underestimate the true effects of treatment [6].

Additional Survey Details

The survey was called the UNIVERSITY-UNIVERSITY Opinion Study, conducted by University #1 and University #2. The survey was conducted by the authors using a panel initially recruited through the mail and then managed using Qualtrics via e-mail, using the e-mail addresses subjects provided us.

The population refers to registered voters in selected neighborhoods in Fresno, Orange County, and Tennessee, as chosen by staff at the partner organizations. Voters were recruited from this population by mail we sent to their household.

The below table shows how the representativeness of those who responded to the survey differ from those mailed an invitation to participate in the survey. These data come from the voter file. Note that no weighting is used in the analysis; the aim of the estimation is to test for the existence of treatment effects within this sample, not to generalize to the population of invited respondents.

Table OA1: Representativeness of Experiment at Each Stage

Sample	Female	Age	AfAm	Latino	Voted 16	Voted 14	Voted 12	Voted 10	Voted 08	N
Starting	0.51	49.08	0.05	0.16	0.74	0.4	0.63	0.45	0.59	217600
Baseline Resp.	0.52	50.03	0.03	0.11	0.85	0.55	0.71	0.55	0.66	7870
Canvassed	0.51	52.23	0.02	0.1	0.89	0.6	0.75	0.61	0.71	2374
1 Wk Resp.	0.52	52.38	0.02	0.09	0.9	0.63	0.75	0.63	0.72	1578
1 Mo Resp.	0.52	52.46	0.02	0.09	0.91	0.63	0.76	0.63	0.73	1508
3-6 Mo Resp.	0.51	52.49	0.02	0.1	0.91	0.63	0.77	0.64	0.73	1384

Outcomes

The survey included dozens of political, social, and cultural questions, only some of which were related to immigration. In our pre-analysis plan, we indicate which items constituted experimental outcomes. Below we list these items and give their full text.

The below items appeared on multiple surveys; the # sign below will be replaced with the survey number in our analysis:

- The baseline survey is survey 0;
- the 1-week survey is survey 1;
- the 1-month survey is survey 2, and;
- the 3-6 month survey is survey 3.

The variable name for each item is written using `in-line code`. For the remainder of the paper, we will refer to these items by their variable names.

Anti-Immigrant Prejudice Index

The first set of questions are five point scales where respondents were asked: "Do you agree or disagree with the below statements about undocumented or illegal immigrants?" Response options were: Strongly agree, Somewhat agree, Neither agree nor disagree, Somewhat disagree, Strongly disagree:

- `t#_imm_prej_living`: "I would have no problem living in areas where undocumented immigrants live."
- `t#_imm_prej_fit`: "Too many undocumented immigrants just don't want to fit into American society."
- `t#_imm_prej_burden`: "Undocumented immigrants are too much of a burden on our communities."
- `t#_imm_prej_crime`: "Undocumented immigrants have already broken the law coming here illegally, so they are more likely to commit other crimes."
- `t#_imm_prej_values`: "Undocumented immigrants hold the same values as me and my family."

Respondents were also asked a feeling thermometer:

- `t#_therm_illegal_immigrant`: Feeling thermometer towards "illegal immigrants".

Respondents were asked a variant of the Bogardus social distance scale question [1]. They were coded as 1 if they answered Relative; 2 for Friend; 3 for Neighbor; 4 for Coworker; 5 for Resident of My State; 6 for None:

- `t#_social_distance_immigrant`: "Below are some groups of people. Look at each of them and say which is the closest relationship you would find acceptable for each group. For example, if you would accept someone from a group living on your street, but not as a close friend, then you would choose neighbors... Undocumented Immigrant". (Note that due to a coding error, this was not included in the t3 survey.)

Anti-Immigrant Policy Index

Respondents were first asked: "Politicians are considering a number of policies about immigration. We want to know what you think. Do you agree or disagree with the statements below?" Response options were: Strongly agree, Somewhat agree, Neither agree nor disagree, Somewhat disagree, Strongly disagree:

- `t#_imm_attorney`: "The government should provide legal aid to all undocumented immigrants who cannot afford their own attorney for legal or courtroom deportation proceedings." This was the question specific to Fresno.
- `t#_imm_police`: "Local police should ask for documentation and automatically turn immigrants over to federal immigration officers when they are found to be in the country illegally." This was the question specific to Orange County.
- `t#_imm_deportall`: "The federal government should work to identify and deport all illegal immigrants, including in the workplace." This was the question specific to Tennessee.
- `t#_imm_daca`: "The federal government should grant legal status to people who were brought to the US illegally as children and who have graduated from a U.S. high school."
- `t#_imm_citizenship`: "The federal government should allow undocumented immigrants currently in the U.S. to become citizens after they have lived, worked, and paid taxes for at least 5 years."
- `t#_imm_compassion`: "Undocumented immigrants deserve compassion and should not live in daily fear of deportation."

Perspective Taking Index

In addition to reducing exclusionary attitudes, we were also interested in measuring whether the canvass increased respondents' abilities to take the perspectives of undocumented immigrants. This was not the primary purpose of the intervention.

Respondents were asked: "Do you agree or disagree with the below statements about undocumented or illegal immigrants?" Response options were: Strongly agree, Somewhat agree, Neither agree nor disagree, Somewhat disagree, Strongly disagree:

- `t#_imm_persp_imagine`: "I can imagine how things look from undocumented immigrants' perspective." This question is abbreviated from [10].
- `t#_imm_persp_difficult`: "I find it difficult to see things from an undocumented immigrants' point of view." This question is abbreviated from [4].

Active Processing Index

We asked the same on respondents' ability to actively process on immigration, which similarly was not the primary purpose of the intervention.

- `t#_imm_actproc_thought`: "I have thought a lot about how we should treat undocumented immigrants in our community."
- `t#_imm_actproc_confident`: "I feel confident in my ability to distinguish good from bad immigration policies." This question is abbreviated from [9].

Outcome Indices

In our pre-analysis plan, we specified that we would combine multiple items into indices to test hypotheses. Combining outcomes into an index increases precision by decreasing survey measurement error and limits the potential for multiple hypothesis testing [3].

The indices, to be described momentarily, are as follows:

- `t#_factor_prej`: An index of outcomes from t#_factor_overall capturing the Anti-Immigrant Prejudice Index.
- `t#_factor_policy`: An index of outcomes from t#_factor_overall capturing the Anti-Immigrant Prejudice Index.
- `t#_factor_overall`: An index of all primary outcomes (i.e., all the items in the prejudice and policy indicies), created to test the omnibus hypothesis that the treatment had any effects.
- `t#_factor_persptake`: An index of outcomes from the Perspective Taking Index.
- `t#_factor_actproc`: An index of outcomes from the Active Processing Index.

In addition to an outcome index at each time period, we also present the results of a pooled outcome index — `tALL_factor_` — taking the average of each outcome index across time periods. If a respondent did not take a particular post-treatment survey, that time period is excluded from the average. Note that this pooled outcome was not pre-specified in our pre-analysis plan. We use this pooled outcome index to increase the precision of our treatment effect estimates through further reductions in measurement error. This pooled outcome is also a useful brief summary of the overall effects. We calculate the pooled outcome using the below Stata code:

```
// Generate factor averages for pooling
foreach factor in prej policy overall actproc persptake {
    egen tALL_factor_`factor' = rowmean(t1_factor_`factor' t2_factor_`factor' t3_factor_`factor')
}
```

Procedure for Combining Outcomes into Indices

We pre-specified that we would create the indices by using factor analysis and rescaling the factors to have mean 0 and standard deviation 1.

We use the below Stata code to generate the factors. Note that we code all indices such that higher values on the indices indicate more tolerance and success of the intervention. If a factor is reverse-coded, we multiply by -1 to adjust for this.

```
factor [VARIABLES USED], fa(1)
predict t#_[FACTOR NAME]_temp
```

```
egen t#_[FACTOR NAME] = std(t#_[FACTOR NAME]_temp) // standardize to mean 0, SD 1
```

Estimation Procedures

Average Treatment Effects

Consistent with our pre-analysis plan, to estimate treatment effects we use ordinary least squares (OLS) regressions with cluster-robust standard errors, clustering on household and also including the pre-treatment covariates from the baseline survey and voter list named in our pre-analysis plan. This procedure and these covariates were pre-specified in advance and produce unbiased estimates of causal effects [6, 3]. Note that there is no reclassification of treatment based on what occurs at the door and we do not exclude any subjects who came to the door; we compare all subjects who came to the door and were pre-assigned to the treatment conversation to all subjects who came to the door and were pre-assigned to the placebo conversation.

Contact Rate

Contact is defined as the voter coming to the door and being identified before the topics of the placebo or immigration begin. Across the three conditions among voters who responded to the baseline survey and were then randomly assigned to an experimental condition, the contact rates were:

- Placebo: 0.31.
- Abbreviated Intervention: 0.29.
- Full Intervention: 0.31.

In the two intervention conditions, we asked the voters to first answer a rating question about immigration. The share of voters who answered this question provides an indication of how often voters actually began the conversation. Conditional on having come to the door, the proportion of voters who provided this first rating is:

- Abbreviated Intervention: 0.73.
- Full Intervention: 0.68.

Across the three conditions, among the voters who came to the door, the canvass completion rates were:

- Placebo: 0.86.
- Abbreviated Intervention: 0.76.
- Full Intervention: 0.66.

Tests of Design Assumptions

Covariate Balance among All Subjects, Compliers, and Reporters

The below tables demonstrate that balance on pre-treatment observable attributes is maintained among the original universe of pre-survey respondents randomized to each group, the sub-sample that was canvassed, and the sub-sample that was both canvassed and successfully re-interviewed. Each table shows the mean value for the covariate under each condition as well as the p-value from a one-way ANOVA test. The first table considers all voters who were randomly assigned after having taken the pre-survey (all subjects); the second table considers all voters who were successfully contacted (compliers); the remaining tables consider all voters who responded to the first through third post-surveys (reporters).

Table OA2: Covariate Balance among Pre-Survey Respondents.

	Placebo	Abbrev Intervention	Full Intervention	p-value
Age	50.02	50.04	50.02	1
Female	0.51	0.52	0.52	0.87
Latino	0.11	0.11	0.11	0.82
Legal Immigrant Feeling Thermometer t0	82.95	83.55	82.44	0.15
Illegal Immigrant Feeling Thermometer t0	47.16	47.23	47.52	0.89
Baseline Factor of Support	0.00	0.00	0.00	0.99
N	2623.00	2623.00	2624.00	-

Table OA3: Covariate Balance among Compliers.

	Placebo	Abbrev Intervention	Full Intervention	p-value
Age	52.07	52.09	52.51	0.83
Female	0.51	0.52	0.50	0.76
Latino	0.12	0.09	0.11	0.16
Legal Immigrant Feeling Thermometer t0	82.43	83.45	82.64	0.6
Illegal Immigrant Feeling Thermometer t0	46.97	47.82	47.40	0.84
Baseline Factor of Support	-0.03	0.03	0.01	0.43
N	814.00	748.00	812.00	-

Table OA4: Covariate Balance among 1st Post-Survey Respondents.

	Placebo	Abbrev Intervention	Full Intervention	p-value
Age	51.42	52.76	52.98	0.25
Female	0.52	0.52	0.53	0.89
Latino	0.11	0.08	0.09	0.16
Legal Immigrant Feeling Thermometer t0	84.06	84.40	82.96	0.48
Illegal Immigrant Feeling Thermometer t0	48.43	48.93	48.66	0.96
Baseline Factor of Support	0.05	0.11	0.08	0.62
N	536.00	499.00	543.00	-

Table OA5: Covariate Balance among 2nd Post-Survey Respondents.

	Placebo	Abbrev Intervention	Full Intervention	p-value
Age	51.48	53.17	52.79	0.23
Female	0.52	0.50	0.52	0.84
Latino	0.09	0.08	0.10	0.64
Legal Immigrant Feeling Thermometer t0	84.33	84.67	83.03	0.4
Illegal Immigrant Feeling Thermometer t0	48.38	49.27	49.55	0.79
Baseline Factor of Support	0.06	0.12	0.09	0.56
N	514.00	486.00	508.00	-

Table OA6: Covariate Balance among 3rd Post-Survey Respondents.

	Placebo	Abbrev Intervention	Full Intervention	p-value
Age	51.83	52.61	53.00	0.53
Female	0.50	0.49	0.54	0.31
Latino	0.11	0.09	0.10	0.6
Legal Immigrant Feeling Thermometer t0	84.15	85.18	82.49	0.12
Illegal Immigrant Feeling Thermometer t0	48.25	48.47	48.79	0.96
Baseline Factor of Support	0.05	0.09	0.09	0.75
N	459.00	448.00	477.00	-

Survey Attrition

An important design assumption is that the treatment does not affect the composition of the individuals who take each follow-up survey [3]. We investigate this by regressing an indicator for responding to a post-treatment survey on indicators of treatment assignment. Across the three survey waves, we find no evidence of differential attrition.

Table OA7: Test for differential attrition

	Effect	SE	t.stat	p
1 Week				
Abbrev	-0.01	0.01	-1.28	0.20
Full	0.00	0.01	0.23	0.82
1 Month				
Abbrev	-0.01	0.01	-0.98	0.33
Full	0.00	0.01	-0.22	0.83
3-6 Months				
Abbrev	0.00	0.01	-0.40	0.69
Full	0.01	0.01	0.65	0.52

Test of Differential Attrition by Covariates

The above subsection demonstrated that there was no average differential attrition; now, we test for whether the treatment caused attrition to differ by covariates (for example, whether it encouraged already-supportive subjects to complete the post-survey but also discouraged unsupportive subjects from doing so) [6]. To test whether attrition patterns are similar by covariates in treatment and placebo, we use a linear regression of whether or not an individual responded to the follow-up survey on treatment, baseline covariates, and treatment-covariate interactions. We then perform a heteroskedasticity-robust F-test of the hypothesis that all the interaction coefficients are zero. This procedure was pre-specified in our pre-analysis plan and is standard practice [6]. Below we report the p-value of this F-test. Based on the results presented in the Table below, there does not appear to be evidence of asymmetrical attrition.

Table OA8: p-value by Survey Wave Test of Differential Attrition by Covariates.

1 Week Survey (t1)	0.57
1 Month Survey (t2)	0.85
3-6 Month Survey (t3)	0.6

Results

Below we report the results in tabular form at each time period and for each outcome measure. In each section, the first table shows the results by the average treatment effect. Each table includes two models: one in which we adjust for the pre-specified pre-treatment covariates to improve precision and a second unadjusted model.

Overall Index of Exclusionary Attitudes

Below we present the ATE on the overall index. Note that we pre-registered a focus on the estimates with covariates (which were also pre-registered) since we expected these to be much more precise; the experimental design was intended to draw significant statistical power from the baseline survey. However, we also present results without covariates for completeness.

Table OA9: ATE effects on overall index

	With Covariates				Without Covariates			
	Effect	SE	t.stat	p	Effect	SE	t.stat	p
1 Week								
Full vs. Placebo	0.090	0.024	3.792	0.000	0.112	0.064	1.764	0.078
Abbrev. vs. Placebo	0.016	0.023	0.664	0.507	0.077	0.064	1.192	0.233
Full vs. Abbrev.	0.073	0.024	3.110	0.002	0.036	0.064	0.557	0.578
1 Month								
Full vs. Placebo	0.068	0.024	2.774	0.006	0.110	0.066	1.687	0.092
Abbrev. vs. Placebo	0.035	0.024	1.487	0.137	0.106	0.065	1.645	0.100
Full vs. Abbrev.	0.033	0.025	1.326	0.185	0.004	0.066	0.061	0.952
3-6 Months								
Full vs. Placebo	0.070	0.027	2.612	0.009	0.115	0.068	1.685	0.092
Abbrev. vs. Placebo	0.018	0.026	0.696	0.487	0.072	0.069	1.040	0.298
Full vs. Abbrev.	0.052	0.026	2.013	0.044	0.043	0.068	0.629	0.529
Pooled								
Full vs. Placebo	0.082	0.021	3.887	0.000	0.093	0.060	1.564	0.118
Abbrev. vs. Placebo	0.022	0.020	1.097	0.273	0.072	0.060	1.200	0.230
Full vs. Abbrev.	0.059	0.021	2.822	0.005	0.021	0.060	0.344	0.731

Prejudice Index

Below we present the ATE on the prejudice index.

Table OA10: ATE effects on prejudice index

	With Covariates				Without Covariates			
	Effect	SE	t.stat	p	Effect	SE	t.stat	p
1 Week								
Full vs. Placebo	0.068	0.027	2.469	0.014	0.092	0.063	1.445	0.149
Abbrev. vs. Placebo	0.004	0.027	0.146	0.884	0.054	0.065	0.825	0.409
Full vs. Abbrev.	0.064	0.027	2.396	0.017	0.038	0.063	0.598	0.550
1 Month								
Full vs. Placebo	0.066	0.028	2.357	0.019	0.109	0.065	1.674	0.094
Abbrev. vs. Placebo	0.040	0.027	1.498	0.134	0.100	0.065	1.537	0.124
Full vs. Abbrev.	0.026	0.028	0.929	0.353	0.009	0.066	0.130	0.897
3-6 Months								
Full vs. Placebo	0.053	0.030	1.769	0.077	0.101	0.068	1.495	0.135
Abbrev. vs. Placebo	0.022	0.030	0.722	0.470	0.074	0.069	1.074	0.283
Full vs. Abbrev.	0.033	0.030	1.107	0.268	0.027	0.068	0.397	0.691
Pooled								
Full vs. Placebo	0.072	0.024	3.020	0.003	0.084	0.059	1.414	0.157
Abbrev. vs. Placebo	0.016	0.023	0.695	0.487	0.058	0.060	0.967	0.334
Full vs. Abbrev.	0.055	0.024	2.355	0.019	0.025	0.060	0.417	0.677

Policy Index

Below we present the ATE on the policy index.

Table OA11: ATE effects on policy index

	With Covariates				Without Covariates			
	Effect	SE	t.stat	p	Effect	SE	t.stat	p
1 Week								
Full vs. Placebo	0.106	0.026	4.122	0.000	0.126	0.064	1.960	0.050
Abbrev. vs. Placebo	0.026	0.026	1.016	0.310	0.096	0.064	1.512	0.131
Full vs. Abbrev.	0.078	0.026	2.947	0.003	0.029	0.064	0.457	0.648
1 Month								
Full vs. Placebo	0.064	0.027	2.393	0.017	0.104	0.066	1.571	0.116
Abbrev. vs. Placebo	0.027	0.026	1.029	0.304	0.106	0.064	1.653	0.098
Full vs. Abbrev.	0.038	0.027	1.394	0.164	-0.002	0.066	-0.035	0.972
3-6 Months								
Full vs. Placebo	0.082	0.030	2.782	0.006	0.122	0.069	1.770	0.077
Abbrev. vs. Placebo	0.013	0.029	0.442	0.659	0.065	0.070	0.936	0.349
Full vs. Abbrev.	0.069	0.028	2.448	0.015	0.056	0.068	0.836	0.403
Pooled								
Full vs. Placebo	0.088	0.023	3.885	0.000	0.097	0.060	1.618	0.106
Abbrev. vs. Placebo	0.029	0.022	1.282	0.200	0.084	0.060	1.401	0.162
Full vs. Abbrev.	0.058	0.022	2.570	0.010	0.013	0.060	0.217	0.828

Perspective Taking and Active Processing Index

Per the pre-analysis plan, as exploratory hypotheses we also asked a perspective-taking index and an active processing-index. Our pooled results are consistent (but not statistically significant) with the full intervention producing more perspective taking than the placebo and abbreviated intervention. Similarly, our results

are consistent (but not statistically significant) with both interventions potentially increasing participants' abilities to think actively about immigration.

With this said, for perspective-taking, upon reflection, the results are difficult to interpret because it is unclear in what direction we might expect effects. For example, a participant who was swayed by the intervention and became less prejudiced towards unauthorized immigrants might say they have difficulty seeing things from an undocumented immigrant's point of view because they now have a much better understanding of the challenges undocumented immigrants face.

Future research should further investigate these mechanisms.

Complier Average Causal Effects

In the pre-analysis plan, we noted that we planned to adjust for the complier average causal effect (CACE) by dividing the average treatment effect (ATE) by the proportion of conversations where the voter answered the first rating question. This CACE assumes that 1) there was no effect of the intervention for the voters who immediately refused to talk, and 2) there are no defiers; that is, no voters only received the intervention if they were assigned to the placebo group yet would not have received it were they actually in the treatment group [6]. Reporting these point estimates would not change the experimental comparison we conduct, but would increase point estimates to account for the measurement error in the treatment indicator.

Although they would be larger, we do not focus on the CACE estimates for two reasons. First, there are multiple reasonable ways to define compliance in this setting. Subjects ended some of the conversations moments after they identified themselves at the door (and hence are included in our sample); others continued further into the conversation but ended the interaction before it was completed. Compliance in this setting is therefore inherently a continuum, and there is no point in the conversation where subjects go from fully non-compliant to fully compliant. Second, there are slightly different compliance rates in the Full and Abbreviated interventions, simply by chance. This complicates comparisons of the Full and Abbreviated interventions when adjusting for compliance. (The Abbreviated Intervention actually had a slightly higher compliance rate, meaning such an adjustment would make our results about the differences between the conditions stronger.)

At the same time, we understand some readers will be interested in a CACE. To compute a CACE, we use a conservative definition of compliance, whether subjects got to the "first rating" part of the conversation where they initially told canvassers how they felt about the policy. Using this definition of compliance, the compliance rate in the Full Intervention condition was 68% and the contact rate in the Abbreviated Intervention condition was 73%. The ITT point estimates should therefore be divided by 1.48 and 1.36 in the Full and Abbreviated interventions, respectively, to compute the CACE. That is, given a true effect on compliers of 1.48 and 1.36 times the size of the ATEs we observed, we would on average estimate ATEs of the magnitude that we did. These imply CACE estimates in the Full and Abbreviated intervention conditions of, respectively, d = 0.122 and d = 0.031.

Heterogeneous Treatment Effects

In this section we present heterogeneous treatment effects by pre-specified subgroups. For each result, we present the conditional average treatment effect, adjusting for covariates, within each subgroup. As the outcome in these analyses we use the index of all the items in the prejudice and policy indicies as measured in the first post-treatment survey.

We pre-specified that we would investigate heterogeneous treatment effects by voter and canvasser traits. We will investigate these by comparing the ATEs within the pre-specified subgroups. The primary groups are:

- By canvasser immigration status: Answered "yes" to "Do you consider yourself to be an immigrant?"
- By t0 having a close friend, colleague, or family member who is undocumented in the baseline survey (yes vs. all other responses).
- By t0 being born outside of the US.

The more exploratory groups are:

- By site.
- By canvasser race: Latino-only vs. White-only vs. Other.
- By canvasser age: 30 and under vs. over 30.
- By t0 party: Democrat vs. Republican vs. Independent/Other.
- By t0 race: Latino-only vs. White-only vs. Asian-only vs. African American-only vs. Other.
- By t0 economic status: Excellent/good personal financial situation vs. Other.
- By t0 education: College educated vs. Other.
- By baseline support: Three separate subgroups of bottom third, middle third, and top third of baseline support factor used to block

Canvasser Immigration Status

Below are results by the immigration status of the *canvasser*. All canvassers completed a demographic survey in which they were asked "Do you consider yourself to be an immigrant?" Responses are coded as 1 for yes and 0 for all other responses.

Table OA12: Heterogeneous treatment effects by Canvasser being an immigrant

	Canvasser being an immigrant = 1				Canvasser being an immigrant = 0			
	Effect	SE	t.stat	p	Effect	SE	t.stat	p
Full	0.12	0.05	2.20	0.03	0.08	0.03	3.12	0.00
Abbrev	0.00	0.05	-0.09	0.93	0.03	0.03	1.15	0.25

Voter Closeness to Unauthorized Immigrants

Below are the results by the closeness of the respondent to unauthorized immigrants. In the baseline survey, we asked "Do you have any close friends, colleagues, or family members who are illegal or undocumented immigrants?". Responses are coded as 1 for yes and 0 for all other responses.

Table OA13: Heterogeneous treatment effects by Voter knows an unauthorized immigrant

	Voter knows an unauthorized immigrant = 1				Voter knows an unauthorized immigrant = 0			
	Effect	SE	t.stat	p	Effect	SE	t.stat	p
Full	0.08	0.06	1.46	0.15	0.08	0.03	3.23	0.00
Abbrev	0.02	0.06	0.36	0.72	0.00	0.02	-0.06	0.95

Voter Immigration Status

Below are the results by the immigration status of the respondent. In the baseline survey, we asked "Which of these statements best describes you?" The provided responses were "I was born in the United States" and "I was born somewhere else". Responses are coded as 1 for being born in the US and 0 for all other responses.

Note that among the respondents to the one week survey who were born outside of the US, there were only 29 people in the full implementation condition, 30 in the abbreviated intervention condition, and 31 in the placebo condition, hence the noisy results.

Table OA14: Heterogeneous treatment effects by Voter born in the US

	Voter born in the US = 1				Voter born in the US = 0			
	Effect	SE	t.stat	p	Effect	SE	t.stat	p
Full	0.09	0.02	3.83	0.00	-0.01	0.10	-0.14	0.89
Abbrev	0.01	0.02	0.40	0.69	0.16	0.09	1.78	0.08

By site

Below are the results by the canvass site. We noted in our pre-analysis plan that this was an exploratory analysis.

Table OA15: Heterogeneous treatment effects by site

	Orange County				Fresno				TN			
	Effect	SE	t.stat	p	Effect	SE	t.stat	p	Effect	SE	t.stat	p
Full	0.04	0.04	0.92	0.36	0.12	0.05	2.58	0.01	0.13	0.04	3.50	0.0
Abbrev	-0.08	0.04	-1.91	0.06	0.10	0.05	2.24	0.02	0.02	0.04	0.52	0.6

By canvasser race

Below are the results by the race of the canvasser. We compare self-identified Latino canvassers to all others. We noted in our pre-analysis plan that this was an exploratory analysis.

Table OA16: Heterogeneous treatment effects by Latino canvasser

	Latino canvasser = 1				Latino canvasser = 0			
	Effect	SE	t.stat	p	Effect	SE	t.stat	p
Full	0.07	0.03	2.14	0.03	0.10	0.03	2.98	0.0
Abbrev	-0.02	0.03	-0.70	0.48	0.06	0.03	1.63	0.1

By canvasser age

Below are the results by the age of the canvasser. We compare canvassers 30 and under vs. over 30. We noted in our pre-analysis plan that this was an exploratory analysis.

Table OA17: Heterogeneous treatment effects by Canvasser under 30

	Canvasser under 30 = 1				Canvasser under 30 = 0			
	Effect	SE	t.stat	p	Effect	SE	t.stat	p
Full	0.07	0.03	2.58	0.01	0.20	0.05	3.77	0.00
Abbrev	-0.01	0.03	-0.34	0.74	0.12	0.05	2.54	0.01

By voter party

Below are the results by the party of the voter. We compare self-identified Democrats to Republicans to Independents (including leaners), as based on responses to the baseline survey. We noted in our pre-analysis plan that this was an exploratory analysis.

Table OA18: Heterogeneous treatment effects by voter party in baseline survey

	Democrat				Republican				Indep/Other			
	Effect	SE	t.stat	p	Effect	SE	t.stat	p	Effect	SE	t.stat	p
Full	0.03	0.04	0.95	0.34	0.14	0.05	2.82	0.00	0.14	0.04	3.28	0.00
Abbrev	0.01	0.04	0.15	0.88	-0.03	0.04	-0.59	0.55	0.08	0.04	1.97	0.05

By voter race

Below are the results by the race of the voter. We compare self-identified Asian to Latino to White voters, as based on responses to the baseline survey. We noted in our pre-analysis plan that this was an exploratory analysis.

Note that among the respondents to the one week survey who were African American, there were only 16 people in the full implementation condition, 12 in the abbreviated intervention condition, and 13 in the placebo condition. Due to the small sample size, we do not include African American in the below table.

Table OA19: Heterogeneous treatment effects by voter race/ethnicity

	Asian				Latino				White			
	Effect	SE	t.stat	p	Effect	SE	t.stat	p	Effect	SE	t.stat	p
Full	0.17	0.11	1.54	0.13	0.22	0.10	2.22	0.03	0.09	0.03	3.51	0.00
Abbrev	-0.07	0.13	-0.50	0.62	0.18	0.09	1.86	0.06	0.01	0.03	0.22	0.83

By voter economic status

Below are the results by the economic status of the voter. In the baseline survey, we asked "How would you rate your own personal financial situation?" Response options were "Excellent", "Good", "Only fair", "Poor", and "Would rather not say". We compare voters who said "Excellent" or "Good" to all other responses. We noted in our pre-analysis plan that this was an exploratory analysis.

Table OA20: Heterogeneous treatment effects by high financial status

	high financial status = 1				high financial status = 0			
	Effect	SE	t.stat	p	Effect	SE	t.stat	p
Full	0.10	0.03	3.32	0.00	0.08	0.04	2.06	0.04
Abbrev	0.04	0.03	1.40	0.16	-0.05	0.04	-1.31	0.19

By voter education

Below are the results by the education status of the voter. We compare voters who self-reported having at least a college degree to all other voters. We noted in our pre-analysis plan that this was an exploratory analysis.

Table OA21: Heterogeneous treatment effects by College educated

	College educated = 1				College educated = 0			
	Effect	SE	t.stat	p	Effect	SE	t.stat	p
Full	0.13	0.03	4.50	0.00	0.04	0.04	0.86	0.39
Abbrev	0.05	0.03	1.92	0.06	-0.04	0.04	-0.96	0.34

By voter baseline support

Below are the results by the baseline support of the voter. This is based on an index combining all of the unauthorized immigration policy and prejudice questions from the baseline survey. We divide this index into terciles and report results for each tercile. We noted in our pre-analysis plan that this was an exploratory analysis.

Table OA22: Heterogeneous treatment effects by immigration support in baseline survey

	Least Supportive				Mid Supportive				Most Suppotive			
	Effect	SE	t.stat	p	Effect	SE	t.stat	p	Effect	SE	t.stat	p
Full	0.09	0.05	1.82	0.07	0.09	0.05	1.75	0.08	0.09	0.03	3.33	0.00
Abbrev	-0.04	0.05	-0.77	0.44	0.04	0.05	0.86	0.39	0.06	0.03	1.88	0.06

Estimates for Dichotomized Policy Items

It may be difficult for readers to interpret the magnitude of an effect presented in terms of standard deviation change. We therefore take two, non-pre-registered approaches to help communicate the substantive size of our estimates.

Strong Support

First, one way to make the results more interpretable is to examine treatment effects on whether participants said they strongly supported the policies asked about in the surveys. This attempts to recreate how participants might vote on each proposal if faced with a ballot measure or was deciding between candidates who differ on their immigration proposals. Note that we did not pre-specify this benchmarking procedure. We use this to illustrate the magnitude of our findings.

We first calculate the share of policies in the follow-up survey that individuals said they strongly supported. To calculate this, we dichotomize five of the policy items (all but `t#_imm_compassion` because this is a general as opposed to specific policy, although the effects on this item are larger), recoding them as 1 if the participant takes the most supportive immigration position in the first follow-up survey (one week post-treatment) and 0 for all other positions in the first follow-up survey. We then take the sum of these positions and compare across the conditions. Note that all estimated effects are intent-to-treat effects, and are not adjusted by the share of individuals who were contacted that actually had some or all of the conversations.

The average share of policies strongly supported in the Placebo group is 0.29.

Comparing the treatment conditions to the placebo, we find that the effect on the share of policies strongly supported of the Abbreviated Intervention condition is -0.005 with a p-value of 0.752; the effect on the share of policies strongly supported of the Full Intervention condition is 0.036 with a p-value of 0.011. The difference between the Full and Abbreviated condition is 0.038 with a p-value of 0.002. These statistics are covariate-adjusted using the same covariates and estimation approach as we pre-specified for the main analysis; although note again that this approach to dichotomizing the items was not pre-specified.

We also report results on the individual dichotimized items, set to 1 if an individual took a strong position on the supportive side of the issue and 0 otherwise.

The results on the individual dichotomized items are as follows:

Table OA23: At max at t1 (ITTs)

	With Covariates			
	Effect	SE	t.stat	p
Attorney				
Full vs. Placebo	0.041	0.021	1.906	0.057
Abbrev. vs. Placebo	0.014	0.022	0.651	0.515
Full vs. Abbrev.	0.024	0.023	1.076	0.282
DACA				
Full vs. Placebo	0.047	0.023	2.063	0.039
Abbrev. vs. Placebo	0.006	0.024	0.238	0.812
Full vs. Abbrev.	0.043	0.023	1.916	0.056
Deport All				
Full vs. Placebo	0.059	0.023	2.614	0.009
Abbrev. vs. Placebo	0.001	0.022	0.031	0.975
Full vs. Abbrev.	0.059	0.023	2.595	0.010
Citizenship				
Full vs. Placebo	0.027	0.024	1.135	0.257
Abbrev. vs. Placebo	-0.029	0.024	-1.230	0.219
Full vs. Abbrev.	0.055	0.024	2.304	0.021
Police				
Full vs. Placebo	0.007	0.022	0.298	0.766
Abbrev. vs. Placebo	-0.014	0.022	-0.618	0.537
Full vs. Abbrev.	0.021	0.023	0.910	0.363
Show Compassion				
Full vs. Placebo	0.061	0.022	2.789	0.005
Abbrev. vs. Placebo	0.024	0.022	1.100	0.271
Full vs. Abbrev.	0.037	0.022	1.648	0.100

Any Support

Second, we also conduct a version of this benchmarking where we dichotomize each variable to record whether participants registered any support (not only strong agreement). These new dichotomized variables are coded to 1 if a participant agreed at all with the policy and 0 otherwise (indicating stated indifference or opposition). In this analysis we again exclude the compassion item (where the effects are the largest but the policy is also not very specific).

The average share of policies supported at all in the Placebo group is 0.55. Comparing the treatment conditions to the placebo, we find that the effect on the share of policies supported at all of the Abbreviated Intervention condition is -0.012 with a p-value of 0.343; the effect on the share of policies strongly supported of the Full Intervention condition is 0.022 with a p-value of 0.058. The difference between the Full and Abbreviated conditions is 0.028 with a p-value of 0.007. These statistics are covariate-adjusted using the same covariates and estimation approach as we pre-specified for the main analysis; although note again that this approach to dichotomizing the items was not pre-specified.

The results on the individual items dichotimized in this manner are as follows:

Table OA24: Agree at all at t1 (ITTs)

	With Covariates			
	Effect	SE	t.stat	p
Attorney				
Full vs. Placebo	0.036	0.023	1.620	0.105
Abbrev. vs. Placebo	0.008	0.023	0.350	0.726
Full vs. Abbrev.	0.028	0.023	1.251	0.211
DACA				
Full vs. Placebo	0.029	0.021	1.392	0.164
Abbrev. vs. Placebo	-0.005	0.022	-0.218	0.827
Full vs. Abbrev.	0.034	0.021	1.635	0.102
Deport All				
Full vs. Placebo	0.027	0.022	1.247	0.213
Abbrev. vs. Placebo	-0.015	0.023	-0.683	0.495
Full vs. Abbrev.	0.044	0.022	2.022	0.043
Citizenship				
Full vs. Placebo	0.015	0.021	0.683	0.495
Abbrev. vs. Placebo	-0.032	0.023	-1.418	0.156
Full vs. Abbrev.	0.045	0.022	1.985	0.047
Police				
Full vs. Placebo	0.002	0.021	0.120	0.904
Abbrev. vs. Placebo	-0.013	0.022	-0.606	0.545
Full vs. Abbrev.	0.017	0.022	0.800	0.424
Show Compassion				
Full vs. Placebo	0.057	0.021	2.740	0.006
Abbrev. vs. Placebo	-0.003	0.022	-0.160	0.873
Full vs. Abbrev.	0.061	0.021	2.906	0.004

Results with Weights

To assess the generalizability of our results, we compare our main results – a sample average treatment effect (SATE) – to an estimate of the population average treatment effect (PATE). As [11] note, "The PATE can only be different from the SATE when two things hold: (1) there is meaningful variation in the treatment impact, and (2) that variation is correlated with the weights... It is important to compare the PATE and SATE estimates. A meaningful discrepancy between them is a signal to look for treatment effect heterogeneity and a flag that weight misspecification could be a real concern. If the estimates do not differ, however, and there is no other evidence of heterogeneity, then extrapolation is less of a concern – and furthermore the SATE is probably a sufficient estimate for the PATE."

To estimate the PATE, we first construct weights of who was canvassed and took the survey relative to the starting universe. We construct these weights using entropy balancing [8] and weight on gender, age, race, and vote history.

Below are results with and without these weights, showing that the estimated SATEs and PATEs are similar. If anything, the estimated PATEs are larger than the estimated SATEs, suggesting that the set of individuals who are canvassed and respond to surveys are perhaps more difficult to persuade than the broader universe.

Note that this analysis was not pre-registered but was prompted by feedback on the draft version of the paper.

Table OA25: ATE Effects on Overall Index with Weights

	Unweighted				Weighted			
	Effect	SE	t.stat	p	Effect	SE	t.stat	p
1 Week								
Full vs. Placebo	0.089	0.024	3.792	0.000	0.112	0.032	3.486	0.001
Abbrev. vs. Placebo	0.016	0.023	0.664	0.507	0.022	0.029	0.733	0.464
1 Month								
Full vs. Placebo	0.068	0.024	2.774	0.006	0.100	0.032	3.130	0.002
Abbrev. vs. Placebo	0.035	0.024	1.487	0.137	0.030	0.027	1.092	0.275
3-6 Months								
Full vs. Placebo	0.070	0.027	2.612	0.009	0.100	0.036	2.816	0.005
Abbrev. vs. Placebo	0.018	0.026	0.696	0.487	0.023	0.032	0.728	0.467

References

[1] Emory Stephen Bogardus. A social distance scale. *Sociology & Social Research*, 1933.

[2] David Broockman and Joshua Kalla. Durably reducing transphobia: A field experiment on door-to-door canvassing. *Science*, 352(6282):220–224, 2016.

[3] David E Broockman, Joshua L Kalla, and Jasjeet S Sekhon. The design of field experiments with survey outcomes: A framework for selecting more efficient, robust, and ethical designs. *Political Analysis*, 25(4):435–464, 2017.

[4] Mark H Davis. Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of personality and social psychology*, 44(1):113, 1983.

[5] Hunter Gehlbach and Maureen E. Brinkworth. The social perspective taking process: Strategies and sources of evidence in taking another's perspective. *Teachers College Record*, 114(1):1–29, 2012.

[6] Alan S Gerber and Donald P Green. *Field experiments: Design, analysis, and interpretation*. WW Norton, 2012.

[7] Alan S Gerber, Donald P Green, Edward H Kaplan, and Holger L Kern. Baseline, placebo, and treatment: Efficient estimation for three-group experiments. *Political Analysis*, 18(3):297–315, 2010.

[8] Jens Hainmueller. Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis*, 20(1):25–46, 2012.

[9] Seth J Hill and Gregory A Huber. On the meaning of survey reports of roll call votes not cast in a legislature. *American Journal of Political Science*, 2018.

[10] Gillian Ku, Cynthia S Wang, and Adam D Galinsky. The promise and perversity of perspective-taking in organizations. *Research in Organizational Behavior*, 35:79–102, 2015.

[11] Luke W. Miratrix, Jasjeet S. Sekhon, Alexander G. Theodoridis, and Luis F. Campos. Worth weighting? how to think about and use weights in survey experiments. *Political Analysis*, 26(3):275–291, 2018.

[12] David W Nickerson. Scalable protocols offer efficient design for field experiments. *Political Analysis*, 13(3):233–252, 2005.

[13] Richard E Petty and John T Cacioppo. *Communication and persuasion: Central and peripheral routes to attitude change*. Springer Series in Social Psychology, 1986.

Online Appendix for Experiments 2 and 3 (Transphobia)

Joshua Kalla[*]
David Broockman[†]

Contents

The full replication code and data that produces this report will be available at https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/8BFYQO. This experiment was preregistered at Evidence in Governance and Politics (EGAP), see https://egap.org/registration/2005.

[*]Yale University, Departments of Political Science and Statistics & Data Science, josh.kalla@yale.edu
[†]University of California, Berkeley, Department of Political Science, dbroockman@berkeley.edu

Survey Recruitment Procedures and Experimental Design

The general survey recruitment procedures and experimental design were identical to Experiment 1 except as otherwise noted below.

Experiments 2 and 3 had the following steps:

- We first attempted to survey 324615 voters in the target universes.
- 17252 voters responded to this survey from 14935 households.
- For Experiment 2, at the household level, these respondents were randomly assigned to receive the Participants' and Video Narratives Condition (1095), the Video Narratives Only Condition (1214), or to a placebo condition about recycling (1176).
- The implementing partners canvassed 384 voters with the Participants' and Video Narratives Condition, 457 voters with the Video Narratives Only Condition, and 493 with a placebo conversation.
- For Experiment 3, we took individuals who either (a) lived outside of the canvass area or (b) were never reached during the canvass phase (e.g., a canvasser never knocked on the door or nobody was home) and randomly assigned them to receive either a phone conversation with the Participants' Narratives Intervention (6892) or a placebo phone conversation about recycling (6898).
- The implementing partners spoke with 1268 voters in the phone intervention and 1369 with placebo calls.
- We then successfully resurveyed 75% of the voters they contacted successfully in a survey one week following the intervention and 73% of the voters they contacted successfully in a survey one month following the intervention.

The below table gives the raw counts for each conditions and other ancillary statistics:

Metric	Canvass Placebo	Canvass Participants' and Video Narratives Condition	Canvass Video Narratives Only	Phone Placebo	Phone Participants' Narratives
Number Assigned	2823	2815	2815	6898	6892
Results: Number Reached	653	556	649	1369	1268
Reached 1st Rating	n/a	294	349	n/a	766
Reached 2nd Rating	n/a	239	304	n/a	527
Reached 3rd Rating	n/a	128	146	n/a	489
Length: 0-2 minutes	208	41	49	1304	454
Length: 2-5 minutes	215	48	88	50	218
Length: 5-10 minutes	17	77	159	9	283
Length: 10-15 minutes	5	92	87	4	189

Metric	Canvass Placebo	Canvass Participants' and Video Narratives Condition	Canvass Video Narratives Only	Phone Placebo	Phone Participants' Narratives
Length: Over 15 minutes	0	60	34	2	70
Voter Shared Story	n/a	202	44	n/a	368
Canvasser Shared Story	n/a	251	69	n/a	477

Outcomes

We conducted two follow-up surveys for Experiments 2 and 3: a first one week after contact and a second one month after contact. We asked voters 15 questions on each survey, 9 about their views towards transgender-inclusive non-discrimination laws and 6 about their tolerance of transgender people.

Working together with the implementation partners, we developed the following outcome measures that appeared on the survey. The survey items were split into two categories: prejudice items and policy items.

During the surveys, we also showed an opposition political ad closely patterned after an ad run by opponents of transgender-inclusive non-discrimination laws in Houston in 2015 alleging that sexual predators would abuse transgender-inclusive non-discrimination laws. This video was included to measure whether treatment effects persist after exposure to counter-arguments [4]. Half of our outcome measures were asked before this video was shown and half were asked after.

Prejudice Items Before Video

- `transprej_comfortwork` I would feel comfortable working closely with a transgender person (a person who was born with a boy's body but now identifies as a woman or a person who was born with a girl's body but now identifies as a man).
- `transprej_moralwrong` Saying you are a gender that is different than the one you were born as is morally wrong.
- `transprej_moralgenderchange` It is morally wrong for someone born with a boy's body to undergo a gender change and live every day as a woman.
- `transprej_friend` I would support a friend choosing to have a sex change.
- `transprej_restroom` It would be wrong to allow a transgender woman (a person who was born with a boy's body but identifies as a woman) to use a woman's restroom or locker room.

Policy Items Before Video

- `transpolicy_LGBTdiscrim` A law in your state that would protect gay and transgender people from discrimination in employment, housing, and public accommodations.
- `transpolicy_lawbathroom` (Starting on second survey.) Our state's nondiscrimination law should allow transgender people to use the restroom that matches the gender they live every day—so a person who lives every day as a woman could use the women's restroom, even if that person was born and raised as a boy.

Prejudice Items After Video

- `transvid_comfortbathroom` I would feel comfortable sharing a bathroom with someone who is transgender.
- `therm_trans` Rating of transgender people on a feeling thermometer.
- `transvid_restroom` It would be wrong to allow a transgender woman (a person who was born with a boy's body but identifies as a woman) to use a woman's restroom or locker room.

Policy Items After Video

- `transvid_LGBTdiscrim` A law in your state that would protect gay and transgender people from discrimination in employment, housing, and public accommodations.
- `transvid_fire` A law protecting transgender people from being fired for being transgender.
- `transvid_school` A law that requires transgender students to use the school bathrooms and locker rooms that match their biological or anatomical sex at birth, rather than the gender they live as every day.
- `transvid_predator` I'm concerned that sexual predators could take advantage of a nondiscrimination law to put women's and children's safety at risk.
- `transvid_teacher` Transgender people should not be allowed to serve as public school teachers.

Computing Indices

We computed our indices using the below code.

```
compute.index.dv <- function(dv.names, survey.wave.boolean.vector){
  responders <- analysis.data[survey.wave.boolean.vector==1,]
  dv.names <- paste0(dv.names, '_scaled')
  index <- rowMeans(responders[,dv.names], na.rm = TRUE)
  index <- scale(index)
  return(index[match(analysis.data$id, responders$id)])
}
```

Estimation Procedures

Consistent with our pre-analysis plan, to estimate treatment effects we use ordinary least squares (OLS) regressions with cluster-robust standard errors, clustering on household and also including the pre-treatment covariates from the baseline survey and voter list named in our pre-analysis plan. This procedure and these covariates were pre-specified in advance and produce unbiased estimates of causal effects [2, 1]. Note that there is no reclassification of treatment based on what occurs at the door and we do not exclude any subjects who came to the door; we compare all subjects who came to the door and were pre-assigned to the treatment conversation to all subjects who came to the door and were pre-assigned to the placebo conversation.

Tests of Design Assumptions

Survey Representativeness

The below tables shows how the representativeness of those who responded to the survey differ from those mailed an invitation to participate in the survey. These data come from the voter file. Note that no weighting is used in the analysis; the aim of the estimation is to test for the existence of treatment effects within this sample, not to generalize to the population of invited respondents.

This first table examines the canvass experiment (Experiment 2).

Table OA26: Representativeness of Canvass Experiment at Each Stage

Sample	Female	Reg. Dem	Reg. Rep	Af-Am	Latino	White	Voted 14	Voted 12	Voted 10	Voted 08	AZ	FL	GA	OH	N
Starting	0.55	0.27	0.3	0.1	0.02	0.82	0.63	0.87	0.64	0.83	0.26	0.25	0.22	0.27	159941
Baseline Resp.	0.54	0.3	0.27	0.05	0.02	0.87	0.77	0.91	0.73	0.85	0.33	0.21	0.26	0.2	8440
Canvassed	0.53	0.34	0.28	0.06	0.02	0.88	0.83	0.93	0.79	0.88	0.27	0.25	0.25	0.23	1858
1 Wk Resp.	0.52	0.33	0.27	0.05	0.02	0.89	0.82	0.93	0.77	0.86	0.25	0.29	0.25	0.21	1044
1 Mo Resp.	0.53	0.34	0.26	0.05	0.02	0.88	0.83	0.92	0.77	0.86	0.25	0.29	0.25	0.21	989

This second table examines the phone experiment (Experiment 3).

Table OA27: Representativeness of Phone Experiment at Each Stage

Sample	Female	Reg. Dem	Reg. Rep	Af-Am	Latino	White	Voted 14	Voted 12	Voted 10	Voted 08	AZ	FL	GA	OH	N
Starting	0.54	0.24	0.36	0.05	0.02	0.88	0.76	0.94	0.78	0.91	0.27	0.22	0.25	0.26	169638
Baseline Resp.	0.52	0.29	0.33	0.04	0.02	0.91	0.84	0.95	0.82	0.91	0.37	0.18	0.23	0.22	13767
Called	0.53	0.32	0.34	0.04	0.01	0.92	0.91	0.97	0.9	0.95	0.32	0.16	0.22	0.3	2637
1 Wk Resp.	0.52	0.33	0.33	0.03	0.01	0.93	0.91	0.98	0.91	0.95	0.32	0.16	0.21	0.31	1943
1 Mo Resp.	0.53	0.33	0.34	0.03	0.01	0.93	0.92	0.98	0.91	0.96	0.32	0.16	0.21	0.31	1897

Covariate Balance among All Subjects, Compliers, and Reporters in Canvass Experiment

The below tables demonstrate that balance on pre-treatment observable attributes is maintained among the original universe of pre-survey respondents randomized to each group, the sub-sample that was canvassed, and the sub-sample that was both canvassed and successfully re-interviewed for the canvass experiment. Each table shows the mean value for the covariate under each condition as well as the p-value from a one-way ANOVA test. The first table considers all voters who were randomly assigned after having taken the pre-survey (all subjects); the second table considers all voters who were successfully contacted (compliers); the remaining tables consider all voters who responded to the first and second post-surveys (reporters).

Table OA28: Covariate Balance among Pre-Survey Respondents, Canvass Experiment

	Placebo	Video Narratives Only Condition	Participants' and Video Narratives Condition	p-value
Registered Democrat	0.29	0.29	0.30	0.7
Registered Republican	0.28	0.27	0.27	0.9
Female	0.52	0.56	0.55	0.05
White	0.88	0.86	0.87	0.06
Transgender People Feeling Thermometer t0	58.35	58.51	58.18	0.89
Donald Trump Feeling Thermometer t0	28.40	28.90	27.81	0.47
Barack Obama Feeling Thermometer t0	54.28	54.41	54.19	0.98
Hillary Clinton Feeling Thermometer t0	42.55	42.99	42.30	0.77
N	2818.00	2811.00	2811.00	-

Table OA29: Covariate Balance among Compliers, Canvass Experiment

	Placebo	Video Narratives Only Condition	Participants' and Video Narratives Condition	p-value
Registered Democrat	0.33	0.34	0.34	0.89
Registered Republican	0.28	0.27	0.28	0.96
Female	0.51	0.55	0.54	0.24
White	0.89	0.86	0.88	0.22
Transgender People Feeling Thermometer t0	58.09	58.15	57.54	0.91
Donald Trump Feeling Thermometer t0	28.01	29.75	29.22	0.64
Barack Obama Feeling Thermometer t0	54.33	54.01	53.22	0.88
Hillary Clinton Feeling Thermometer t0	44.09	44.43	42.56	0.65
N	653.00	649.00	556.00	-

Table OA30: Covariate Balance among 1st Post-Survey Respondents, Canvass Experiment

	Placebo	Video Narratives Only Condition	Participants' and Video Narratives Condition	p-value
Registered Democrat	0.32	0.35	0.32	0.67
Registered Republican	0.28	0.26	0.28	0.82
Female	0.52	0.53	0.53	0.92
White	0.90	0.87	0.90	0.28
Transgender People Feeling Thermometer t0	59.76	59.79	58.18	0.66
Donald Trump Feeling Thermometer t0	28.21	25.84	29.47	0.36
Barack Obama Feeling Thermometer t0	54.43	57.26	53.70	0.45
Hillary Clinton Feeling Thermometer t0	43.10	46.21	42.16	0.32
N	387.00	352.00	305.00	-

Table OA31: Covariate Balance among 2nd Post-Survey Respondents, Canvass Experiment

	Placebo	Video Narratives Only Condition	Participants' and Video Narratives Condition	p-value
Registered Democrat	0.34	0.34	0.36	0.86
Registered Republican	0.26	0.26	0.27	0.98
Female	0.52	0.55	0.52	0.69
White	0.90	0.87	0.88	0.43
Transgender People Feeling Thermometer t0	59.67	60.66	59.68	0.85
Donald Trump Feeling Thermometer t0	27.89	25.71	28.24	0.58
Barack Obama Feeling Thermometer t0	55.72	57.52	55.69	0.78
Hillary Clinton Feeling Thermometer t0	44.23	46.66	43.54	0.52
N	369.00	334.00	286.00	-

Covariate Balance among All Subjects, Compliers, and Reporters in Phone Experiment

The below tables demonstrate that balance on pre-treatment observable attributes is maintained among the original universe of pre-survey respondents randomized to each group, the sub-sample that was called, and the sub-sample that was both called and successfully re-interviewed for the phone experiment. Each table shows the mean value for the covariate under each condition as well as the p-value from a t-test, given there are only two groups. The first table considers all voters who were randomly assigned after having taken the pre-survey (all subjects); the second table considers all voters who were successfully contacted (compliers); the remaining tables consider all voters who responded to the first and second post-surveys (reporters).

Table OA32: Covariate Balance among Pre-Survey Respondents, Phone Experiment

	Placebo	Phone Intervention	p-value
Registered Democrat	0.28	0.29	0.22
Registered Republican	0.33	0.33	0.96
Female	0.52	0.52	0.71
White	0.90	0.91	0.72
Transgender People Feeling Thermometer t0	56.24	56.21	0.95
Donald Trump Feeling Thermometer t0	31.51	31.32	0.74
Barack Obama Feeling Thermometer t0	50.08	49.97	0.87
Hillary Clinton Feeling Thermometer t0	40.23	40.03	0.75
N	6888.00	6879.00	-

Survey Attrition in Canvass Experiment

An important design assumption is that the treatment does not affect the composition of the individuals who take each follow-up survey [1]. We investigate this by regressing an indicator for responding to a post-treatment survey on indicators of treatment assignment. Across the two survey waves, we find slight

Table OA33: Covariate Balance among Compliers, Phone Experiment

	Placebo	Phone Intervention	p-value
Registered Democrat	0.31	0.34	0.15
Registered Republican	0.34	0.33	0.51
Female	0.52	0.55	0.16
White	0.92	0.92	0.57
Transgender People Feeling Thermometer t0	54.38	56.53	0.03
Donald Trump Feeling Thermometer t0	33.98	32.38	0.25
Barack Obama Feeling Thermometer t0	48.05	50.96	0.06
Hillary Clinton Feeling Thermometer t0	40.04	41.92	0.2
N	1369.00	1268.00	-

Table OA34: Covariate Balance among 1st Post-Survey Respondents, Phone Experiment

	Placebo	Phone Intervention	p-value
Registered Democrat	0.31	0.34	0.14
Registered Republican	0.34	0.33	0.75
Female	0.51	0.53	0.38
White	0.94	0.93	0.49
Transgender People Feeling Thermometer t0	55.38	57.15	0.13
Donald Trump Feeling Thermometer t0	31.58	30.60	0.53
Barack Obama Feeling Thermometer t0	49.59	52.35	0.13
Hillary Clinton Feeling Thermometer t0	41.05	43.00	0.25
N	1014.00	929.00	-

Table OA35: Covariate Balance among 2nd Post-Survey Respondents, Phone Experiment

	Placebo	Phone Intervention	p-value
Registered Democrat	0.32	0.34	0.31
Registered Republican	0.34	0.34	0.77
Female	0.52	0.54	0.21
White	0.94	0.93	0.46
Transgender People Feeling Thermometer t0	55.02	57.35	0.05
Donald Trump Feeling Thermometer t0	31.84	30.42	0.37
Barack Obama Feeling Thermometer t0	49.43	52.66	0.08
Hillary Clinton Feeling Thermometer t0	40.48	43.13	0.12
N	994.00	903.00	-

evidence of differential attrition. In the below sections, we show that pre-treatment covariates do not predict this slight differential attrition.

Table OA36: Test for differential attrition, Canvass Experiment

	Effect	SE	t.stat	p
1 Week				
Video Narratives Only Condition	-0.05	0.03	-1.83	0.07
Participants' and Video Narratives Condition	-0.04	0.03	-1.54	0.12
1 Month				
Video Narratives Only Condition	-0.05	0.03	-1.82	0.07
Participants' and Video Narratives Condition	-0.05	0.03	-1.76	0.08

Survey Attrition in Phone Experiment

In the below table we look at the phone experiment and again, across the two survey waves, we find no evidence of differential attrition.

Table OA37: Test for differential attrition, Phone Experiment

	Effect	SE	t.stat	p
1 Week				
Treat	-0.01	0.02	-0.47	0.64
1 Month				
Treat	-0.01	0.02	-0.80	0.43

Test of Differential Attrition by Covariates in Canvass Experiment

The above subsection demonstrated that there was no average differential attrition; now, we test for whether the treatment caused attrition to differ by covariates (for example, whether it encouraged already-supportive subjects to complete the post-survey but also discouraged unsupportive subjects from doing so) [2]. To test whether attrition patterns are similar by covariates in treatment and placebo, we use a linear regression of whether or not an individual responded to the follow-up survey on treatment, baseline covariates, and treatment-covariate interactions. We then perform a heteroskedasticity-robust F-test of the hypothesis that all the interaction coefficients are zero. Below we report the p-value of this F-test. Based on the results presented below, there does not appear to be evidence of asymmetrical attrition.

Table OA38: p-value by Survey Wave Test of Differential Attrition by Covariates, Canvass Experiment

1 Week Survey (t1)	0.68
1 Month Survey (t2)	0.38

Test of Differential Attrition by Covariates in Phone Experiment

Below we present the same test for the phone experiment.

Table OA39: p-value by Survey Wave Test of Differential Attrition by Covariates, Phone Experiment

1 Week Survey (t1)	0.75
1 Month Survey (t2)	0.53

Canvass Results (Experiment 2)

Effects on Overall Index

First, we show the effects on an overall index that combines all the outcomes above together. Overall, we see statistically significant effects from all types of conversations and that these effects persist for at least one month.

Table OA40: ATE effects on overall index

	With Covariates				Without Covariates			
	Effect	SE	t.stat	p	Effect	SE	t.stat	p
1 Week								
Participants' and Video Narratives Condition vs. Placebo	0.093	0.027	3.494	0.000	0.021	0.079	0.267	0.790
Video Narratives Only Condition vs. Placebo	0.094	0.024	3.969	0.000	0.047	0.075	0.619	0.536
Participants' and Video Narratives vs. Video Narratives Only	-0.002	0.028	-0.062	0.951	-0.025	0.082	-0.308	0.758
1 Month								
Participants' and Video Narratives Condition vs. Placebo	0.080	0.026	3.049	0.002	0.064	0.080	0.804	0.422
Video Narratives Only Condition vs. Placebo	0.064	0.025	2.513	0.012	0.041	0.077	0.531	0.596
Participants' and Video Narratives vs. Video Narratives Only	0.016	0.030	0.538	0.591	0.023	0.083	0.276	0.783
Pooled								
Participants' and Video Narratives Condition vs. Placebo	0.079	0.024	3.346	0.001	0.043	0.076	0.564	0.573
Video Narratives Only Condition vs. Placebo	0.079	0.022	3.572	0.000	0.031	0.073	0.426	0.670
Participants' and Video Narratives vs. Video Narratives Only	-0.001	0.026	-0.022	0.983	0.012	0.080	0.148	0.882

Effects on Policy Index

Next, we show the effects on the policy index.

Table OA41: ATE effects on policy index

	With Covariates				Without Covariates			
	Effect	SE	t.stat	p	Effect	SE	t.stat	p
1 Week								
Participants' and Video Narratives Condition vs. Placebo	0.089	0.034	2.660	0.008	0.029	0.078	0.367	0.714
Video Narratives Only Condition vs. Placebo	0.072	0.031	2.359	0.018	0.036	0.074	0.488	0.626
Participants' and Video Narratives vs. Video Narratives Only	0.018	0.035	0.501	0.617	-0.007	0.082	-0.090	0.928
1 Month								
Participants' and Video Narratives Condition vs. Placebo	0.070	0.032	2.217	0.027	0.057	0.079	0.721	0.471
Video Narratives Only Condition vs. Placebo	0.054	0.030	1.807	0.071	0.028	0.077	0.361	0.718
Participants' and Video Narratives vs. Video Narratives Only	0.018	0.035	0.498	0.618	0.030	0.084	0.351	0.726
Pooled								
Participants' and Video Narratives Condition vs. Placebo	0.073	0.029	2.522	0.012	0.041	0.076	0.546	0.585
Video Narratives Only Condition vs. Placebo	0.065	0.028	2.369	0.018	0.020	0.072	0.282	0.778
Participants' and Video Narratives vs. Video Narratives Only	0.010	0.031	0.303	0.762	0.021	0.080	0.260	0.795

Effects on Prejudice Index

Next, we show the effects on the prejudice index.

Table OA42: ATE effects on prejudice index

	With Covariates				Without Covariates			
	Effect	SE	t.stat	p	Effect	SE	t.stat	p
1 Week								
Participants' and Video Narratives Condition vs. Placebo	0.091	0.028	3.308	0.001	0.015	0.080	0.184	0.854
Video Narratives Only Condition vs. Placebo	0.104	0.025	4.213	0.000	0.052	0.076	0.683	0.495
Participants' and Video Narratives vs. Video Narratives Only	-0.015	0.029	-0.535	0.593	-0.037	0.082	-0.448	0.654
1 Month								
Participants' and Video Narratives Condition vs. Placebo	0.087	0.029	3.010	0.003	0.068	0.080	0.844	0.399
Video Narratives Only Condition vs. Placebo	0.070	0.027	2.629	0.009	0.053	0.078	0.687	0.492
Participants' and Video Narratives vs. Video Narratives Only	0.014	0.031	0.435	0.664	0.014	0.082	0.176	0.860
Pooled								
Participants' and Video Narratives Condition vs. Placebo	0.083	0.025	3.335	0.001	0.045	0.077	0.582	0.561
Video Narratives Only Condition vs. Placebo	0.089	0.023	3.925	0.000	0.041	0.074	0.558	0.577
Participants' and Video Narratives vs. Video Narratives Only	-0.008	0.027	-0.314	0.753	0.004	0.080	0.048	0.962

Effects on Pre-Video Index

As described above, halfway through the surveys we showed an opposition political ad closely patterned after an ad run by opponents of transgender-inclusive non-discrimination laws in Houston in 2015 alleging that sexual predators would abuse transgender-inclusive non-discrimination laws. This video was included to measure whether treatment effects persist after exposure to counter-arguments [4].

Next, we show the effects on the index of questions asked before the opposition video was shown. This is a mix of prejudice and policy questions.

Table OA43: ATE effects on pre-video index

	With Covariates				Without Covariates			
	Effect	SE	t.stat	p	Effect	SE	t.stat	p
1 Week								
Participants' and Video Narratives Condition vs. Placebo	0.066	0.028	2.362	0.018	-0.019	0.080	-0.240	0.810
Video Narratives Only Condition vs. Placebo	0.074	0.027	2.758	0.006	0.015	0.075	0.198	0.843
Participants' and Video Narratives vs. Video Narratives Only	-0.009	0.030	-0.310	0.757	-0.034	0.082	-0.414	0.679
1 Month								
Participants' and Video Narratives Condition vs. Placebo	0.084	0.029	2.920	0.004	0.060	0.080	0.746	0.456
Video Narratives Only Condition vs. Placebo	0.072	0.028	2.627	0.009	0.041	0.077	0.527	0.598
Participants' and Video Narratives vs. Video Narratives Only	0.011	0.031	0.352	0.725	0.019	0.082	0.229	0.819
Pooled								
Participants' and Video Narratives Condition vs. Placebo	0.070	0.024	2.894	0.004	0.022	0.077	0.292	0.770
Video Narratives Only Condition vs. Placebo	0.075	0.023	3.198	0.001	0.018	0.073	0.249	0.804
Participants' and Video Narratives vs. Video Narratives Only	-0.006	0.026	-0.216	0.829	0.004	0.080	0.055	0.956

Effects on Post-Video Index

Next, we show the effects on the index of questions asked after the opposition video was shown. This is a mix of prejudice and policy questions.

Table OA44: ATE effects on post-video index

	With Covariates				Without Covariates			
	Effect	SE	t.stat	p	Effect	SE	t.stat	p
1 Week								
Participants' and Video Narratives Condition vs. Placebo	0.110	0.031	3.529	0.000	0.052	0.078	0.663	0.507
Video Narratives Only Condition vs. Placebo	0.106	0.027	3.855	0.000	0.070	0.076	0.920	0.358
Participants' and Video Narratives vs. Video Narratives Only	0.004	0.033	0.129	0.897	-0.017	0.082	-0.212	0.832
1 Month								
Participants' and Video Narratives Condition vs. Placebo	0.075	0.030	2.490	0.013	0.067	0.080	0.835	0.404
Video Narratives Only Condition vs. Placebo	0.056	0.028	1.973	0.049	0.040	0.077	0.520	0.603
Participants' and Video Narratives vs. Video Narratives Only	0.019	0.034	0.548	0.584	0.026	0.084	0.314	0.753
Pooled								
Participants' and Video Narratives Condition vs. Placebo	0.083	0.027	3.062	0.002	0.058	0.076	0.762	0.446
Video Narratives Only Condition vs. Placebo	0.082	0.026	3.191	0.002	0.041	0.074	0.559	0.576
Participants' and Video Narratives vs. Video Narratives Only	0.003	0.030	0.094	0.925	0.017	0.080	0.208	0.835

Phone Results (Experiment 3)

Effects on Overall Index

First, we show the effects on an overall index that combines all the outcomes above together. Overall, we see statistically significant effects from all types of conversations and that these effects persist for at least one month.

Table OA45: ATE effects on overall index

	With Covariates				Without Covariates			
	Effect	SE	t.stat	p	Effect	SE	t.stat	p
1 Week								
Participants' Narratives Condition vs. Placebo	0.046	0.015	3.115	0.002	0.080	0.046	1.738	0.082
1 Month								
Participants' Narratives Condition vs. Placebo	0.044	0.016	2.780	0.006	0.114	0.047	2.438	0.015
Pooled								
Participants' Narratives Condition vs. Placebo	0.045	0.014	3.247	0.001	0.090	0.044	2.038	0.042

Effects on Policy Index

Next, we show the effects on the policy index.

Table OA46: ATE effects on policy index

	With Covariates				Without Covariates			
	Effect	SE	t.stat	p	Effect	SE	t.stat	p
1 Week								
Participants' Narratives Condition vs. Placebo	0.034	0.019	1.829	0.068	0.073	0.046	1.568	0.117
1 Month								
Participants' Narratives Condition vs. Placebo	0.029	0.018	1.592	0.112	0.098	0.047	2.098	0.036
Pooled								
Participants' Narratives Condition vs. Placebo	0.033	0.016	1.999	0.046	0.080	0.044	1.815	0.070

Effects on Prejudice Index

Next, we show the effects on the prejudice index.

Table OA47: ATE effects on prejudice index

	With Covariates				Without Covariates			
	Effect	SE	t.stat	p	Effect	SE	t.stat	p
1 Week								
Participants' Narratives Condition vs. Placebo	0.052	0.016	3.196	0.001	0.081	0.046	1.770	0.077
1 Month								
Participants' Narratives Condition vs. Placebo	0.058	0.018	3.308	0.001	0.125	0.046	2.689	0.007
Pooled								
Participants' Narratives Condition vs. Placebo	0.054	0.015	3.601	0.000	0.095	0.044	2.166	0.030

Effects on Pre-Video Index

As described above, halfway through the surveys we showed an opposition political ad closely patterned after an ad run by opponents of transgender-inclusive non-discrimination laws in Houston in 2015 alleging that sexual predators would abuse transgender-inclusive non-discrimination laws. This video was included to measure whether treatment effects persist after exposure to counter-arguments [4].

Next, we show the effects on the index of questions asked before the opposition video was shown. This is a mix of prejudice and policy questions.

Table OA48: ATE effects on pre-video index

	With Covariates				Without Covariates			
	Effect	SE	t.stat	p	Effect	SE	t.stat	p
1 Week								
Participants' Narratives Condition vs. Placebo	0.036	0.017	2.091	0.037	0.060	0.046	1.298	0.195
1 Month								
Participants' Narratives Condition vs. Placebo	0.050	0.018	2.800	0.005	0.111	0.047	2.379	0.017
Pooled								
Participants' Narratives Condition vs. Placebo	0.046	0.015	2.989	0.003	0.083	0.044	1.875	0.061

Effects on Post-Video Index

Next, we show the effects on the index of questions asked after the opposition video was shown. This is a mix of prejudice and policy questions.

Table OA49: ATE effects on post-video index

	With Covariates				Without Covariates			
	Effect	SE	t.stat	p	Effect	SE	t.stat	p
1 Week								
Participants' Narratives Condition vs. Placebo	0.052	0.017	3.076	0.002	0.092	0.046	2.006	0.045
1 Month								
Participants' Narratives Condition vs. Placebo	0.039	0.018	2.219	0.027	0.112	0.047	2.402	0.016
Pooled								
Participants' Narratives Condition vs. Placebo	0.043	0.015	2.824	0.005	0.093	0.044	2.110	0.035

Other Results

Results with Weights

As described in the SM for Experiment 1, to assess the generalizability of our results, we compare our main results – a sample average treatment effect (SATE) – to an estimate of the population average treatment effect (PATE). To estimate the PATE, we first construct weights of who was canvassed and took the survey relative to the starting universe. We construct these weights using entropy balancing [3] and weight on gender, age, race, party registration, and vote history.

Below are results with and without these weights, showing that the estimated SATEs and PATEs are similar. If anything, the estimated PATE is usually larger than the SATE, suggesting that the set of individuals who are canvassed and respond to surveys are perhaps more difficult to persuade than the broader universe.

Note that this analysis was not pre-registered but was prompted by feedback on the draft version of the paper.

Canvass

Table OA50: ATE Effects on Overall Index with Weights, Canvass Experiment

	Unweighted				Weighted			
	Effect	SE	t.stat	p	Effect	SE	t.stat	p
1 Week								
Participants' and Video Narratives Condition vs. Placebo	0.093	0.027	3.494	0.000	0.111	0.029	3.821	0.000
Video Narratives Only Condition vs. Placebo	0.094	0.024	3.969	0.000	0.093	0.029	3.201	0.001
1 Month								
Participants' and Video Narratives Condition vs. Placebo	0.080	0.026	3.049	0.002	0.109	0.028	3.859	0.000
Video Narratives Only Condition vs. Placebo	0.064	0.025	2.513	0.012	0.060	0.030	1.978	0.048

Phone

Table OA51: ATE Effects on Overall Index with Weights, Phone Experiment

	Unweighted				Weighted			
	Effect	SE	t.stat	p	Effect	SE	t.stat	p
1 Week								
Participants' Narratives Condition vs. Placebo	0.046	0.015	3.115	0.002	0.038	0.018	2.148	0.032
1 Month								
Participants' Narratives Condition vs. Placebo	0.044	0.016	2.781	0.005	0.047	0.020	2.327	0.020

Subgroups

In our pre-analysis plan, we specified that we would examine treatment effect heterogeneity by:

- Whether the ATE of canvassing is different for canvassers who identify as transgender or gender non-conforming than for all other canvassers,
- Whether the ATE of canvassing is different for canvassers who are perceived as gender conforming,
- Whether the ATE of canvassing is different for paid vs. volunteer staff, and
- Whether the ATE of canvassing and calling varies by the political knowledge of the participant.

While we did not specify this in the pre-analysis plan, we will also investigate treatment effect heterogeneity by the party identification of the participant.

Note that we only collected the demographics of canvassers, not the callers. We collected the canvasser demographics via survey were they described their gender identity and how they anticipate others perceive their gender identity.

For this subgroup analysis, we present ATE results on the overall index in the 1 week survey.

By canvasser gender identity

In the canvass, 384 conversations and one week surveys were completed by canvassers who self-identify as transgender or gender non-conforming; 596 by canvassers who self-identify as cis-gender; and 64 by canvassers for whom we are missing data.

Table OA52: Heterogeneous treatment effects by gender identity

	Transgender or Gender Non-Conforming				Cisgender				Missing Data			
	Effect	SE	t.stat	p	Effect	SE	t.stat	p	Effect	SE	t.stat	p
Participants' and Video Narratives	0.11	0.04	2.89	0.0	0.07	0.04	1.98	0.05	0.19	0.15	1.24	0.22
Video Narratives Only	0.07	0.04	1.63	0.1	0.09	0.03	3.07	0.00	0.24	0.11	2.24	0.03

By canvasser gender perception

In the canvass, 550 conversations and one week surveys were completed by canvassers who self-identify as being perceived as gender conforming; 430 by canvassers who self-identify as being perceived as gender non-conforming; and 64 by canvassers for whom we are missing data.

Table OA53: Heterogeneous treatment effects by gender identity perception

	Perceived as gender conforming				Perceived as gender non-conforming				Missing Data			
	Effect	SE	t.stat	p	Effect	SE	t.stat	p	Effect	SE	t.stat	p
Participants' and Video Narratives	0.08	0.04	2.02	0.04	0.10	0.04	2.42	0.02	0.19	0.15	1.24	0.22
Video Narratives Only	0.06	0.03	1.87	0.06	0.11	0.04	2.93	0.00	0.24	0.11	2.24	0.03

By canvasser volunteer status

In the canvass, 422 conversations and one week surveys were completed by paid canvassers; 558 by volunteer canvassers; and 64 by canvassers for whom we are missing data.

Table OA54: Heterogeneous treatment effects by canvasser volunteer status

	Paid canvasser				Volunteer canvasser				Missing Data			
	Effect	SE	t.stat	p	Effect	SE	t.stat	p	Effect	SE	t.stat	p
Participants' and Video Narratives	0.13	0.04	2.85	0.00	0.07	0.03	2.08	0.04	0.19	0.15	1.24	0.22
Video Narratives Only	0.08	0.04	1.95	0.05	0.10	0.03	2.94	0.00	0.24	0.11	2.24	0.03

By participant political knowledge (canvass)

In the first post-treatment survey, we asked five political knowledge questions on the length of a presidential term, the length of a Senate term, the purpose of Medicare, the Chief Justice of the Supreme Court, and on which program the US spends the least. We then coded individuals into how many questions they answered correctly out of 5.

For sample size considerations, we group together all respondents who answered 0, 1, or 2 questions correctly.

Table OA55: Heterogeneous treatment effects by political knowledge

	5/5				4/5				3/5				<2/5			
	Effect	SE	t.stat	p	Effect	SE	t.stat	p	Effect	SE	t.stat	p	Effect	SE	t.stat	p
Participants' and Video Narratives	0.08	0.05	1.61	0.11	0.11	0.04	2.82	0	0.11	0.07	1.44	0.15	0.12	0.06	1.92	0.06
Video Narratives Only	0.07	0.05	1.60	0.11	0.13	0.04	3.20	0	0.09	0.05	1.72	0.09	0.13	0.07	2.02	0.04

By participant party identification (canvass)

In the canvass, 407 conversations and one week surveys were completed by voters who self-identified in the baseline survey as Democrats; 295 by self-identified Republicans; and 342 by self-identified Independents (including party leaners).

Table OA56: Heterogeneous treatment effects by voter party identification

	Democrats				Republicans				Independents			
	Effect	SE	t.stat	p	Effect	SE	t.stat	p	Effect	SE	t.stat	p
Participants' and Video Narratives	0.12	0.04	2.94	0	0.15	0.06	2.57	0.01	0.04	0.04	0.90	0.37
Video Narratives Only	0.12	0.03	3.45	0	0.13	0.05	2.43	0.02	0.06	0.04	1.33	0.18

By participant political knowledge (phone)

For sample size considerations, we group together all respondents who answered 0, 1, or 2 questions correctly.

Table OA57: Heterogeneous treatment effects by political knowledge

	5/5				4/5				3/5				<2/5			
	Effect	SE	t.stat	p	Effect	SE	t.stat	p	Effect	SE	t.stat	p	Effect	SE	t.stat	p
Phone	0.05	0.03	2	0.05	0.02	0.02	0.64	0.52	0.07	0.03	2.32	0.02	0.07	0.04	1.64	0.1

By participant party identification (phone)

In the phone, 727 conversations and one week surveys were completed by voters who self-identified in the baseline survey as Democrats; 605 by self-identified Republicans; and 611 by self-identified Independents (including party leaners).

Table OA58: Heterogeneous treatment effects by voter party identification

	Democrats				Republicans				Independents			
	Effect	SE	t.stat	p	Effect	SE	t.stat	p	Effect	SE	t.stat	p
Phone	0.04	0.02	2.1	0.04	0.03	0.03	0.91	0.36	0.06	0.03	2.25	0.02

Estimates for Dichotomized Policy Items

It may be difficult for readers to interpret the magnitude of an effect presented in terms of standard deviation change. We therefore take two, non-pre-registered approaches to help communicate the substantive size of our estimates.

Strong Support - Canvass (Experiment 2)

First, one way to make the results more interpretable is to examine treatment effects on whether participants said they strongly supported the policies asked about in the surveys. This attempts to recreate how participants might vote on each proposal if faced with a ballot measure or was deciding between candidates who differ on their transgender non-discrimination proposals. Note that we did not pre-specify this benchmarking

procedure. We use this to illustrate the magnitude of our findings. In particular, we report results on the individual dichotimized items from the first post-treatment survey, set to 1 if an individual took a strong position on the supportive side of the issue and 0 otherwise. For these analyes for Experiment 2, we combine the two treatments for simplicity and statistical power.

The results on the individual dichotomized items are as follows:

Table OA59: At max at t1 (ITTs)

Variable Name	Effect	SE	t.stat	p
Non-Discrimination Law (Before Video)	0.025	0.022	1.143	0.253
Non-Discrimination Law (After Video)	0.081	0.022	3.588	0.000
Protect From Firing For Being Trans	0.062	0.022	2.758	0.006
Must Use Bathroom Or Locker Matching Sex At Birth (Reverse Coded)	0.033	0.025	1.309	0.191
Concerned About Sex Predators and Non-Discrim Law (Reverse Coded)	0.023	0.021	1.081	0.280
Trans People Should Not Be Public School Teachers (Reverse Coded)	0.014	0.025	0.572	0.567

Any Support - Canvass (Experiment 2)

Second, we also conduct a version of this benchmarking where we dichotomize each variable to record whether participants registered any support (not only strong agreement). These new dichotomized variables are coded to 1 if a participant agreed at all with the policy and 0 otherwise (indicating stated indifference or opposition). The results on the individual items dichotimized in this manner are as follows:

Table OA60: Agreement at all at t1 (ITTs)

Variable Name	Effect	SE	t.stat	p
Non-Discrimination Law (Before Video)	-0.010	0.018	-0.550	0.583
Non-Discrimination Law (After Video)	0.027	0.021	1.281	0.201
Protect From Firing For Being Trans	0.025	0.019	1.316	0.188
Must Use Bathroom Or Locker Matching Sex At Birth (Reverse Coded)	0.024	0.025	0.942	0.347
Concerned About Sex Predators and Non-Discrim Law (Reverse Coded)	0.063	0.020	3.180	0.002
Trans People Should Not Be Public School Teachers (Reverse Coded)	0.042	0.022	1.920	0.055

Strong Support - Phone (Experiment 3)

The results on the individual dichotomized items when dichotomized to capture strong support only for the phone experiment (Experiment 3) are as follows:

Table OA61: At max at t1 (ITTs)

Variable Name	Effect	SE	t.stat	p
Non-Discrimination Law (Before Video)	0.023	0.016	1.468	0.142
Non-Discrimination Law (After Video)	0.020	0.016	1.256	0.209
Protect From Firing For Being Trans	0.043	0.016	2.650	0.008
Must Use Bathroom Or Locker Matching Sex At Birth (Reverse Coded)	-0.003	0.017	-0.169	0.866
Concerned About Sex Predators and Non-Discrim Law (Reverse Coded)	-0.011	0.014	-0.740	0.460
Trans People Should Not Be Public School Teachers (Reverse Coded)	-0.010	0.018	-0.574	0.566

Any Support - Phone (Experiment 3)

The results on the individual dichotomized items when dichotomized to capture any support for the phone experiment (Experiment 3) are as follows:

Table OA62: Agreement at all at t1 (ITTs)

Variable Name	Effect	SE	t.stat	p
Non-Discrimination Law (Before Video)	-0.032	0.014	-2.331	0.020
Non-Discrimination Law (After Video)	-0.014	0.015	-0.929	0.353
Protect From Firing For Being Trans	0.000	0.015	-0.013	0.990
Must Use Bathroom Or Locker Matching Sex At Birth (Reverse Coded)	0.019	0.018	1.065	0.287
Concerned About Sex Predators and Non-Discrim Law (Reverse Coded)	0.012	0.014	0.853	0.394
Trans People Should Not Be Public School Teachers (Reverse Coded)	0.015	0.016	0.931	0.352

References

[1] David E Broockman, Joshua L Kalla, and Jasjeet S Sekhon. The design of field experiments with survey outcomes: A framework for selecting more efficient, robust, and ethical designs. *Political Analysis*, 25(4):435–464, 2017.

[2] Alan S Gerber and Donald P Green. *Field experiments: Design, analysis, and interpretation*. WW Norton, 2012.

[3] Jens Hainmueller. Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis*, 20(1):25–46, 2012.

[4] Richard E. Petty, Curtis P. Haugtvedt, and Stephen M. Smith. Elaboration as a determinant of attitude strength: Creating attitudes that are persistent, resistant, and predictive of behavior. In R.E. Petty and J.A. Krosnick, editors, *Attitude strength: Antecedents and consequences*, pages 93–130. Lawrence Erlbaum Associates, Inc., 1995.