Online Appendix: Personalizing Moral Reframing in Interpersonal Conversation

Joshua L. Kalla* Adam Seth Levine† David E. Broockman‡

Contents

Intervention Details **3**
- Training . . . 3
- Canvasser Demographics . . . 3
- Intervention Procedure . . . 3
- Placebo Procedure . . . 5
- Scripts . . . 5

Survey Recruitment Procedures and Experimental Design **9**
- Baseline Survey . . . 9
- Random Assignment of Households . . . 9
- Random Assignment of Turfs . . . 9
- Placebo Design for Delivering Intervention . . . 10
- Follow-Up Surveys . . . 11
- Additional Survey Details . . . 11

Outcomes **11**
- Outcome Indices . . . 13
- Procedure for Combining Outcomes into Indices . . . 13

Estimation Procedures **14**
- Average Treatment Effects . . . 14
- Contact Rate . . . 14
- Complier Average Causal Effects . . . 14

Tests of Design Assumptions **14**
- Covariate Balance among All Subjects, Compliers, and Reporters . . . 14
- Survey Attrition . . . 16
- Test of Differential Attrition by Covariates . . . 16

Results **17**
- Stigma Index . . . 17
- Policy Index . . . 17
- Action-Taking Index . . . 18
- Planned Parenthood Thermometer . . . 18
- Heterogeneous Treatment Effects . . . 19
- Benchmarking Results . . . 21

*Yale University, josh.kalla@yale.edu
†Johns Hopkins University
‡University of California, Berkeley

Results with Weights 22
Comparison with Broockman, Kalla, and Sekhon (2017) 23

The full replication code and data that produces this report will be available at the JOP Data Archive on Dataverse.

This experiment was pre-registered at https://osf.io/rft8w/.

Intervention Details

Training

Before beginning the experiment we worked with the partner organization for 8 months (spread over two summers) to develop and pilot test the intervention and develop the training. The canvassers were all volunteers. They received training when they first started, including being paired with a more experienced canvasser for a period of time. Throughout the program they received ongoing training and feedback, as well as opportunities for reflection. The training focused on providing canvassers with the skills to listen to and ask questions of voters in a non-judgmental context. They learned about and practiced relationship-building techniques such as using openers (i.e. asking questions rather than lecturing), responsiveness (e.g. responding directly to what the voter said with relevant follow-up questions), and being affirming (i.e. respecting, and not challenging, what voters share). The training also focused on making them comfortable with having conversations about a sensitive topic with strangers. These trainings were all generic trainings on how to have a high-quality canvass conversation.

On abortion in particular, the canvassers received trainings in two specific skills. First, given our focus on personalized moral reframing, the training paid particular attention to Moral Foundations Theory and how canvassers could identify the moral values underlying the beliefs and experiences that voters share along with how to present new arguments that are tied to those values. For example, canvassers completed a values worksheet where they would read a particular statement a voter might say related to a certain moral frame and then discuss ways to respond to this. This worksheet focused on the values of care, fairness, loyalty, authority, sanctity, and liberty.

Second, canvassers also received extensive training related to abortion facts. For example, canvassers received facts about how many women who choose to have an abortion are already parents.

Overall, trainings involved a combination of role play and viewing video of past canvass conversations.

Canvasser Demographics

The canvassers for this project were volunteer canvassers recruited by the local partner organization. Below we summarize the demographics of the canvassers:

- The average canvasser was 37 years old.
- 77% were female.
- 92% were white.
- 25% had children.
- 26% reported having had an abortion.

Intervention Procedure

The canvassers were trained to follow the below procedure when approaching homes when subjects were in the treatment condition. 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 connect on values and 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.

The intervention had the following six steps.

1. Establish contact Determine if the voter is home. The canvasser knocks on the door and says, "Hi, are you (voter's name)?" If the subject identifies him/herself, then 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. Collect initial ratings and follow-up questions to surface experiences, beliefs, and values The canvassers began the intervention by engaging in a series of strategies to elicit participants' opinions in a non-judgmental manner. First, they informed voters that they were talking with folks in their neighborhood about abortion, and acknowledged that it is not a topic that people frequently talk about. Canvassers then asked voters to report their opinion on how permissible abortion should be on a scale from 0-10 and then also asked them to reflect upon and explain their response. Canvassers were trained to probe responses to learn more about the reasons underlying these opinions, all while not indicating that they were pleased or displeased with any particular answer. Instead, the goal was for them to appear genuinely interested in hearing the voter reflect on the questions.

3. Surface voters' beliefs and attitudes Afterwards, the canvasser asked a number of follow-up questions to further understand the experiences that shaped voters' views on abortion, and in particular their previous experiences related to abortion and pregnancy. Canvassers would respond directly and ask for more information, and also ask voters to reflect upon how they/others felt during these experiences. Canvassers also listened carefully for the underlying moral values that the voter was expressing. For instance, voters might cite moral values such as loyalty (e.g., standing by one's friends even if you don't see eye to eye), fairness (e.g., the desire to treat others fairly), and care (e.g., caring about one's family and wanting to do right by them). Overall, the goal was to encourage effortful reflection and to build rapport, and also for the canvasser to learn about the underlying values tied to the voter's beliefs and attitudes.

4. Canvasser identifies values he/she and voter share Next, the canvassers shared their own story, including their own personal experiences that informed their beliefs about abortion as well as underlying values as well. While they did that, they also explicitly acknowledged and named the values they share in common with the voter (e.g., "I know you also believe it's important to stand by your friends, even if you don't see eye to eye"). To facilitate these exchanges, canvassers received pre-canvass trainings on Moral Foundations Theory and moral reframing and prepared multiple stories they could tell, depending on which value matched their own experiences and appeared most salient to the voter. Identifying shared values was important for the canvasser to further establish trust and credibility, as someone who could speak on behalf of those values.

5. Canvasser engages in moral reframing to make the case Next, canvassers would shift the conversation from personal beliefs and values toward the topic of what the experience should be like for a woman who has decided to end a pregnancy. They would first ask voters to reflect on that question. Afterwards, the canvasser would share his/her main argument: that safe, legal abortion should be available for all women and that women should be supported in their decisions and not be judged at all. While doing so, the canvasser would tie this argument to the values shared in common with the voter. This was personalized moral reframing at work. For example, if the conversation had surfaced a shared moral value of fairness, then the canvasser might say "It is only fair that women are supported in their decisions..." Afterwards the canvasser would ask the voter to reflect on what was said, once more taking the opportunity to affirm shared values.

6. Wrap-up Finally, the canvasser would circle back to the question they began with: how permissible abortion should be on a scale from 0-10. The canvasser would ask them to reflect upon and explain any changes. Then they would also ask voters who identified themselves as supporters of abortion access to contact their elected officials to state that support.

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, are you [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 and coded as "canvassed" in our data. Note that this first step is identical in the placebo and treatment condition.
2. **Placebo begins.** The canvasser would then deliver the placebo survey.
3. **Conversation ends.** The canvasser thanks the subject and leaves.

Scripts

Below are the scripts for the Treatment and Placebo conditions.

Part 1: *Introduction: First Ratings*

Hi, are you ____(voter's name)? **[CONFIRM CORRECT VOTER CAME TO DOOR]**
My name is _____ and I'm with the Maine Action Team. We're talking with voters in your neighborhood today about abortion. Abortion is an issue that can bring out strong feelings for people and is not something that we talk about every day.

Today we are asking folks, on a scale of 0 to 10, where 0 means there is no reason a woman should be able to get an abortion, 10 means women should be able to get an abortion for any reason, and in the middle means you think some reasons are okay and some are not, **where would you put yourself?**

0	1	2	3	4	5	6	7	8	9	10
No Reason				**← ← Some Reasons →→**						**Any Reason**

Why does that number feel right to you? What's on either side of the issue for you?
What are some reasons that you think it is ok? Not ok?

Are you from Maine? How long have you lived here? What brought you to Maine?

Turn Over to continue

Part 5: Final Ratings and Explanation of Why

Now that we've been talking about this, on that same scale from before, where 0 means there is no reason a woman should be able to get an abortion, 10 means women should be able to get an abortion for any reason, and in the middle means you think some reasons are okay and some are not, **where would you put yourself?**

0	1	2	3	4	5	6	7	8	9	10
No Reason				**← ← Some Reasons →→**						**Any Reason**

Why does that feel right for you?
[if number increased] ***What changed your mind?***
Why do you feel differently now than you did before?

<u>ASK EVERYONE:</u> **What's the best number to reach you at?** [write it on the walk list]

Part 6 [if 9 or 10 ONLY]: Take action

The Maine Action Team is part of Planned Parenthood here in Maine. Do you support the work of Planned Parenthood? *[if yes]* What is your best email address? [write it on the walk list]

So we are asking folks like yourself to call Senator Susan Collins and let her know you support safe and legal abortion and ask her to stand with Planned Parenthood. Will you call Senator Collins right now and leave her a message?

[if yes] Great! I will dial the phone number and hand it to you. Just leave a message, your name, address, and tell her you support safe and legal abortion and ask her to stand with Planned Parenthood. (*202-224-2523*)

[if no] No worries. Thank you for talking with me today.

Start Time: _____ End Time: _____

Part 2: Surface experiences, beliefs, and values

I'm a volunteer with the Maine Action Team, a part of Planned Parenthood, and as I mentioned, I am talking to people in your neighborhood about abortion. I'm curious, what shaped your views on abortion?

Tell me more about that…
How has that informed your beliefs?
What are some times in your life when you talked about abortion?

Have you known anyone who has had an abortion or considered having one?

[*if none*] **What about someone who found themselves pregnant and wasn't planning to be? Or didn't want to be? or had a pregnancy scare?**

Tell me more about that…
How did that make you feel? How did she feel?
Do you think she was supported? Did you provide support?

(consider the value(s) the voter is expressing, validate and connect or challenge)

3: Draw the line between values and abortion. Explore voter's reactions.

When I think about what has shaped my own beliefs about abortion, I think about…
Share your story. Include personal details.

Tap into your shared values with voter:
you also really care about your family and want to do right by them.
making your own decisions is also really important to you.
You also really try and treat others fairly.
You also believe it's important to stand by your friends, even if you don't see eye to eye.

Pause for reaction

What do you think? How does that make you feel? Tell me more…

4: Surface and address contradictions. Build resolution. Make your case.

I'm curious, what do you think the experience should be like for a woman who has decided to end a pregnancy?

Supported? Safe? Comfortable? Dignified? Should they have to walk through protesters? Should it be hard to get an abortion?

I really enjoyed talking with you today and you being so willing to share. I think we found somethings we agree on in our conversation. It's why I am out here talking to voters- because I believe that safe, legal abortion should be available for all women. I believe women should be supported in their decisions and not be judged. After all, everyone's path in life looks a little different.

Tell me more…Listen and affirm common ground. Wrap up and transition back to final rating on the front.

Start Time: _____ End Time: _____

PLACEBO Medicaid Script
FINAL

Hi, are you ____(voter's name)?

[CONFIRM CORRECT VOTER CAME TO DOOR]

My name is _____ and I'm with the Maine Action Team. We are conducting a brief survey today with voters in your neighborhood about Medicaid expansion.

Intro
Last year, Maine voters passed a law to expand access to healthcare for **70,000 more Mainers.** Have you heard about that?

ID – Expansion
This law expands access to Medicaid insurance to individuals earning less than $17,000 a year.

This new law provides critical health care and makes Maine's economy healthier. It will help fund rural hospitals and community clinics and create thousands of new jobs for nurses, doctors and health care workers. The governor and the legislature have not done what they need to in order to implement this law.

Do you support the governor and the legislature acting to implement the law to expand Medicaid?

If yes or no: *Ok and* what is the best phone number to reach you at? [write on walk list]

Thank you for your time.

Have a good day.

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 Maine. To recruit voters to these surveys, the partner organization 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. When selecting voters whose contact information to provide us with, the organization selected broadly, only removing voters with an out-of-state mailing address, a seasonal mailing address, a small number of voters who it expected were staunch opponents or supporters of abortion based on proprietary analytical models or prior contact with the organization. Voters living in Portland or South Portland were removed. The study was limited to voters living in Androscoggin, Cumberland and York Counties in Maine. We invited these voters to the baseline survey by mail. The survey was called the "Maine 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 70842 households that contained individual logins for 112010 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.

We also conducted a second baseline survey to further improve precision. If subjects did not respond to this second baseline survey, mean values were imputed.

Random Assignment of Households

3355 voters completed at least one baseline survey and provided a valid email address. 7 voters who completed this survey were removed from the experiment at this point because they had moved or had requested to be removed from the Maine Opinion Study. This left 3348 voters to be randomized. We randomly assigned half to treatment and half to placebo. 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 [1] and follows the best practices for field experiments with survey outcomes [2].

The household-level clustered random assignment took place within blocks of two households. These blocks were formed by matching households on household size, whether or not the household responded to the second baseline survey, and on household-level-average factor of abortion views. 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 [2].

Random Assignment of Turfs

On the day of each canvass, groups of households were formed into "turfs" by the staff at the partner organization. "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 gender.

Placebo Design for Delivering Intervention

Canvassers attempted to have a conversation about an unrelated issue (Medicaid expansion) with voters in the placebo group. This placebo-controlled experimental design [7] is common in studies of door-to-door canvassing interventions and field experiments more generally [2]. Nickerson [7] 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. [4] 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 Medicaid expansion began instead). This design dramatically improves the precision of door-to-door canvassing experiments [2].

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 [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 abortion or Medicaid expansion 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 [7, 2].

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 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 [3].

Additional Survey Details

The survey was called the "Maine Opinion Study, a study being conducted at University #1 and University #2 of registered voters from across Maine." 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 Maine, as chosen by staff at the partner organization. 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	White	N
Starting	0.51	46.87	0.99	112010
Baseline Resp.	0.54	47.62	0.99	3355
2nd Baseline Resp.	0.54	47.63	0.99	2220
Canvassed	0.52	50.26	1	1034
1 Wk Resp.	0.52	50.83	1	680
1 Mo Resp.	0.52	50.67	1	679
3 Mo Resp.	0.52	50.7	0.99	612

Outcomes

The survey included dozens of political, social, and cultural questions, only some of which were related to abortion 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 second baseline survey is survey 1;
- the 1-week survey is survey 2;
- the 1-month survey is survey 3, and;

- the 3 month survey is survey 4.

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.

Abortion Policy/Legal Scale The first set of policy/legal questions asked "Do you think abortion should be banned or allowed in the following circumstances?" Respondents were given the response options "Ban abortion" or "Allow abortion".

- `t#_allow_6weeks`: If a woman has already been pregnant for six weeks.
- `t#_allow_trimester`: If a woman has already been pregnant for more than one trimester (more than 12 weeks of pregnancy).
- `t#_allow_no_birthcontrol`: If a woman was not using birth control.
- `t#_allow_partner`: If a woman's partner disagrees with her decision.
- `t#_allow_already`: If a woman has already had an abortion before.
- `t#_allow_anyreason`: If a woman wants an abortion for any reason.

The second set of policy/legal questions asked "How much do you agree or disagree with the following statements regarding abortion?" Respondents were given a 7-point Agree-Disagree scale with a neutral midpoint of "Neither agree nor disagree".

- `t#_law_appointment`: Maine law should require a woman to attend multiple appointments over several days in order to get an abortion.
- `t#_law_private_insur`: Maine law should prohibit private health insurance from covering abortions.
- `t#_law_counsel`: Maine law should require women to receive counseling that advises childbirth over abortion.
- `t#_law_public_insur`: Maine law should require that publicly-funded health insurance for poor people covers the cost of abortions.

A final policy/legal questions asked "Maine is considering a number of ballot measures during the next few years. Below are some of the ballot measures under consideration. Would you favor or oppose each ballot measure?" Respondents were given a 5-point Favor-Oppose scale with a neutral midpoint of "Neither Favor nor Oppose".

- `t#_ballot_abortion`: Prohibiting doctors from performing abortions on women who are beyond 20 weeks of pregnancy.

Abortion Social/Stigma/Moral Scale The first set of social/stigma/moral questions asked "How much do you agree or disagree with the following statements regarding abortion?" Respondents were given a 7-point Agree-Disagree scale with a neutral midpoint of "Neither agree nor disagree".

- `t#_abortion_something_wrong`: Women who have had abortions have done something wrong.
- `t#_abortion_badly`: Women who have had abortions should feel badly about themselves.
- `t#_abortion_consider`: If a woman does not want to be pregnant, she should consider an abortion.
- `t#_abortion_birthcontrol`: Given the availability of modern birth control, women who have abortions are just irresponsible.
- `t#_abortion_nothing_wrong`: There's nothing wrong with having an abortion.

In addition, respondents were asked a 0-100 feeling thermometer for "Women who have had abortions." This is coded as `t#_therm_women_abort`.

Abortion Action-Taking Scale The abortion action-taking questions asked "These days people are busy and often don't have time to do many of the things they would like to. Suppose in the next month someone asked you to engage in the following activities. How likely would you be to say yes?" Respondents were given a 5-point response scale including "Extremely likely", "Very likely", "Somewhat likely", "A little likely", and "Not likely at all".

- t#_act_volunteer_clinic: Volunteer for an organization that supports women who need abortions.
- t#_act_accompany: Accompany a friend or family member to an abortion appointment.
- t#_act_protest_clinic: Protest outside an abortion clinic to show my opposition to abortion.
- t#_act_congress_support: Call a politician to express my support for legal and available abortion.
- t#_act_congress_opposition: Call a politician to express my opposition to legal and available abortion.

Additional Measure: Planend Parenthood Feeling Thermometer We also asked a feeling thermometer towards Planned Parenthood (coded as t#_therm_pp). This was not included in any of the above scales because it did not fit any of the above categories. We asked about Planned Parenthood because they are a commonly-known group in American politics that supports access to abortion.

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 [2].

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

- t#_factor_policy: An index of outcomes from Abortion Policy/Legal Scale.
- t#_factor_stigma: An index of outcomes from Abortion Social/Stigma/Moral Scale.
- t#_factor_action: An index of outcomes from Abortion Action-Taking Scale.

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 policy stigma action {
    egen tALL_factor_`factor' = rowmean(t2_factor_`factor' t3_factor_`factor' t4_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
```

Note that all indices are coded so that positive values reflect greater support for abortion.

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 [3, 2]. 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 treatment begin. Across the two conditions among voters who responded to the baseline survey and were then randomly assigned to an experimental condition, the contact rates were:

- Placebo: 0.32.
- Treatment: 0.3.

On average, the conversations lasted:

- Placebo: 1.99 minutes (standard deviation of 1.9 minutes).
- Treatment: 11.83 minutes (standard deviation of 8.39 minutes). Among those who at least provided the first rating, the conversations lasted an average of 13.07 minutes and had a standard deviation of 7.8 minutes.

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 [3]. 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.

We do not report the CACE estimates due to space constraints, but note that these can be computed by taking the ATE estimates and multiplying by approximately 1.12, given the contact rate of approximately 89%. That is, given a true effect on compliers of 1.12 times the size of the ATEs we observed, we would on average estimate ATEs of the magnitude that we did.

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 (measured in the baseline survey before treatment) 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	Treat	p-value
Baseline Abortion Factor	0.00	0.00	0.96
Party ID	-0.49	-0.51	0.76
Planned Parenthood Therm	59.38	58.47	0.45
Women Who Had Abortions Therm	58.80	58.89	0.92
African American Therm	79.60	80.32	0.36
Female	0.54	0.54	0.98
Age	47.62	47.58	0.93
2nd Baseline Respondent	0.67	0.66	0.81
N	1669.00	1679.00	-

Table OA3: Covariate Balance among Compliers.

	Placebo	Treat	p-value
Baseline Abortion Factor	0.02	0.07	0.44
Party ID	-0.53	-0.51	0.87
Planned Parenthood Therm	58.44	59.42	0.65
Women Who Had Abortions Therm	58.11	60.38	0.12
African American Therm	78.87	80.95	0.15
Female	0.52	0.52	0.93
Age	50.67	49.82	0.26
2nd Baseline Respondent	0.67	0.68	0.78
N	532.00	502.00	-

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

	Placebo	Treat	p-value
Baseline Abortion Factor	0.08	0.17	0.25
Party ID	-0.49	-0.44	0.7
Planned Parenthood Therm	58.57	62.28	0.17
Women Who Had Abortions Therm	58.88	61.59	0.14
African American Therm	79.59	81.32	0.33
Female	0.53	0.51	0.57
Age	51.43	50.16	0.16
2nd Baseline Respondent	0.85	0.85	1
N	357.00	323.00	-

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

	Placebo	Treat	p-value
Baseline Abortion Factor	0.06	0.12	0.39
Party ID	-0.45	-0.43	0.92
Planned Parenthood Therm	58.46	61.87	0.21
Women Who Had Abortions Therm	59.09	61.19	0.25
African American Therm	79.75	80.88	0.52
Female	0.54	0.50	0.36
Age	50.97	50.36	0.51
2nd Baseline Respondent	0.86	0.85	0.66
N	350.00	329.00	-

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

	Placebo	Treat	p-value
Baseline Abortion Factor	0.10	0.19	0.26
Party ID	-0.39	-0.44	0.75
Planned Parenthood Therm	59.73	63.72	0.16
Women Who Had Abortions Therm	58.47	62.16	0.05
African American Therm	79.43	82.18	0.14
Female	0.54	0.51	0.49
Age	51.28	50.08	0.22
2nd Baseline Respondent	0.87	0.85	0.5
N	316.00	296.00	-

Survey Attrition

An important design assumption is that the treatment does not affect the composition of the individuals who take each follow-up survey [2]. 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				
Treat	-0.02	0.01	-1.55	0.12
1 Month				
Treat	-0.01	0.01	-0.99	0.32
3 Months				
Treat	-0.01	0.01	-0.98	0.33

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) [3]. 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 [3]. 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 (t2)	0.29
1 Month Survey (t3)	0.33
3 Month Survey (t4)	0.55

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. 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.

Stigma Index

Below we present the ATE on the stigma index.

Table OA9: ATE effects on stigma index

	With Covariates				Without Covariates			
	Effect	SE	t.stat	p	Effect	SE	t.stat	p
1 Week								
Treat vs. Placebo	0.032	0.038	0.847	0.397	0.134	0.078	1.709	0.088
1 Month								
Treat vs. Placebo	0.022	0.037	0.595	0.552	0.111	0.078	1.419	0.156
3 Months								
Treat vs. Placebo	0.045	0.038	1.194	0.233	0.154	0.082	1.875	0.061
Pooled								
Treat vs. Placebo	0.044	0.029	1.481	0.139	0.131	0.072	1.817	0.070

Policy Index

Below we present the ATE on the policy index.

Table OA10: ATE effects on policy index

	With Covariates				Without Covariates			
	Effect	SE	t.stat	p	Effect	SE	t.stat	p
1 Week								
Treat vs. Placebo	0.061	0.032	1.933	0.054	0.112	0.079	1.417	0.157
1 Month								
Treat vs. Placebo	0.040	0.033	1.216	0.224	0.080	0.079	1.014	0.311
3 Months								
Treat vs. Placebo	0.032	0.036	0.878	0.380	0.087	0.083	1.051	0.294
Pooled								
Treat vs. Placebo	0.039	0.028	1.395	0.164	0.077	0.073	1.056	0.291

Action-Taking Index

Below we present the ATE on the action-taking index.

Table OA11: ATE effects on action-taking index

	With Covariates				Without Covariates			
	Effect	SE	t.stat	p	Effect	SE	t.stat	p
1 Week								
Treat vs. Placebo	0.186	0.051	3.663	0.000	0.291	0.078	3.718	0.000
1 Month								
Treat vs. Placebo	0.105	0.052	2.034	0.042	0.190	0.078	2.431	0.015
3 Months								
Treat vs. Placebo	0.080	0.054	1.501	0.134	0.209	0.082	2.544	0.011
Pooled								
Treat vs. Placebo	0.105	0.042	2.499	0.013	0.213	0.069	3.084	0.002

Planned Parenthood Thermometer

Below we present the ATE on the Planned Parenthood feeling thermometer. Like all other items, this outcome is standardized to have mean 0 and standard deviation 1.

Table OA12: ATE effects on Planned Parenthood feeling thermometer

	With Covariates				Without Covariates			
	Effect	SE	t.stat	p	Effect	SE	t.stat	p
1 Week								
Treat vs. Placebo	0.113	0.035	3.243	0.001	0.187	0.078	2.397	0.017
1 Month								
Treat vs. Placebo	0.067	0.037	1.806	0.071	0.139	0.078	1.789	0.074
3 Months								
Treat vs. Placebo	0.064	0.038	1.694	0.091	0.162	0.082	1.974	0.049
Pooled								
Treat vs. Placebo	0.089	0.030	3.000	0.003	0.147	0.073	2.010	0.045

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 pooled average of the post-treatment surveys.

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

- Voters high in political knowledge (`t0_knowledge_scale`) will be less persuadable.
- Voters in the middle of the abortion scale (`t0_abortion_factor`) will be more persuadable.

For ease of presentation, we analyze these subgroup effects by splitting the scale into thirds and examining treatment effects within each third.

While we did not pre-specify these heterogeneous treatment effects, we also examine how treatment effects vary by canvasser demographics (canvassers completed a demographic survey to gather this data):

- Whether the canvasser is female or not.
- Whether the canvasser has children or not.
- Whether the canvasser has previously had an abortion or not.

Canvasser Gender Below are results whether the canvasser self-reported being female or not. 77% of subjects were canvassed by women.

Table OA13: Heterogeneous treatment effects by Canvasser is female

	Canvasser is female = 1				Canvasser is female = 0			
	Effect	SE	t.stat	p	Effect	SE	t.stat	p
Effect on Action-Taking	0.11	0.05	2.22	0.03	0.08	0.11	0.77	0.44
Effect on Stigma	0.05	0.04	1.41	0.16	0.00	0.06	-0.01	0.99
Effect on Policy	0.04	0.03	1.30	0.19	0.04	0.06	0.77	0.44

Canvasser has children Below are results whether the canvasser self-reported having children or not. 25% of subjects were canvassed by canvassers with children.

Table OA14: Heterogeneous treatment effects by Canvasser has children

	Canvasser has children = 1				Canvasser has children = 0			
	Effect	SE	t.stat	p	Effect	SE	t.stat	p
Effect on Action-Taking	0.00	0.09	-0.02	0.98	0.14	0.05	2.79	0.00
Effect on Stigma	0.06	0.06	1.03	0.30	0.06	0.04	1.64	0.10
Effect on Policy	0.03	0.06	0.48	0.64	0.05	0.03	1.75	0.08

Canvasser has had an abortion Below are results whether the canvasser self-reported having had an abortion or not. 26% of subjects were canvassed by a canvasser who self-reported having had an abortion.

Table OA15: Heterogeneous treatment effects by Canvasser has had an abortion

	Canvasser has had an abortion = 1				Canvasser has had an abortion = 0			
	Effect	SE	t.stat	p	Effect	SE	t.stat	p
Effect on Action-Taking	0.07	0.09	0.78	0.44	0.13	0.05	2.65	0.01
Effect on Stigma	0.01	0.06	0.09	0.93	0.06	0.04	1.56	0.12
Effect on Policy	0.02	0.05	0.47	0.64	0.04	0.03	1.16	0.25

By voter political knowledge In the baseline survey, we asked four political knowledge questions. Roughly one-third correctly answered 0, 1, or 2 of these questions, one-third correctly answered 3, and one-third answered four. The questions were:

- Do you happen to remember which party controls the United States Senate – that is, which party has a majority of members in the United States Senate?
- How about the Maine House of Representatives? Do you happen to know which party has the most members in the Maine House of Representatives right now?
- Do you happen to remember who is the current US Attorney General?
- Do you happen to remember what industry the Dodd-Frank Act regulates?

Table OA16: Heterogeneous treatment effects by political knowledge

	0, 1, 2 correct				3 correct				4 correct			
	Effect	SE	t.stat	p	Effect	SE	t.stat	p	Effect	SE	t.stat	p
Effect on Action-Taking	0.22	0.07	3.04	0.00	0.08	0.08	0.97	0.34	0.06	0.07	0.86	0.39
Effect on Stigma	0.11	0.06	1.90	0.06	0.03	0.05	0.59	0.56	0.03	0.05	0.64	0.52
Effect on Policy	0.10	0.05	1.97	0.05	0.04	0.05	0.91	0.36	0.02	0.06	0.37	0.72

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 abortion policy and prejudice questions from the baseline survey. We divide this index into terciles and report results for each tercile.

To better contextualize these terciles, we provide breakdowns of baseline attitudes across each of the terciles:

Tercile	% Democratic at Baseline	Average "Women who have had abortions" Feeling Thermometer at Baseline	% "Not at all likely" to volunteer for an organization that supports women who need abortions
Least Supportive	3%	46	95%
Mid Supportive	11%	58	74%
Most Supportive	24%	75	44%

Table OA18: Heterogeneous treatment effects by abortion 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
Effect on Action-Taking	0.06	0.04	1.31	0.19	0.13	0.07	1.95	0.05	0.09	0.09	1.00	0.32
Effect on Stigma	0.02	0.07	0.33	0.74	0.06	0.06	0.95	0.34	0.01	0.03	0.38	0.71
Effect on Policy	-0.04	0.07	-0.59	0.55	0.04	0.05	0.74	0.46	0.04	0.02	1.60	0.11

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. This was not mentioned in our pre-analysis plan, but given the partisan nature of abortion as an issue, we choose to report it as an exploratory result.

Table OA19: 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
Effect on Action-Taking	0.24	0.16	1.48	0.14	0.12	0.06	1.93	0.06	0.06	0.06	0.94	0.35
Effect on Stigma	0.06	0.08	0.66	0.51	-0.04	0.06	-0.68	0.50	0.10	0.04	2.44	0.01
Effect on Policy	0.05	0.05	1.04	0.30	0.06	0.05	1.16	0.25	0.03	0.04	0.64	0.52

By voter gender Below are results whether the voter is listed on the voter file as being female or not. This was not mentioned in our pre-analysis plan, but was suggested by a reviewer.

Table OA20: Heterogeneous treatment effects by Voter is female

	Voter is female = 1				Voter is female = 0			
	Effect	SE	t.stat	p	Effect	SE	t.stat	p
Effect on Action-Taking	0.13	0.06	1.95	0.05	0.10	0.05	1.95	0.05
Effect on Stigma	0.06	0.04	1.36	0.17	0.03	0.04	0.71	0.48
Effect on Policy	0.05	0.04	1.32	0.19	0.02	0.04	0.59	0.56

By voter education Below are results whether the voter self-reported having a college degree or not. This was not mentioned in our pre-analysis plan, but was suggested by a reviewer.

Table OA21: Heterogeneous treatment effects by Voter is college educated

	Voter is college educated = 1				Voter is college educated = 0			
	Effect	SE	t.stat	p	Effect	SE	t.stat	p
Effect on Action-Taking	0.12	0.06	2.12	0.04	0.12	0.09	1.35	0.18
Effect on Stigma	0.04	0.04	0.92	0.36	0.06	0.06	1.10	0.27
Effect on Policy	0.04	0.04	1.01	0.31	0.09	0.06	1.58	0.12

Benchmarking Results

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** (provided the highest level of agreement) the policies, sentiments, and actions asked about in the surveys. This attempts to recreate how participants might vote on each proposal if faced with a ballot measure or faced with a decision. Note that we did not pre-specify this benchmarking procedure. We use this to illustrate the magnitude of our findings. Note that, in parentheses, P, S, and A refer to the policy, stigma, and action-taking items. An asterisk is used when the predicted effect should be negative.

The results on the individual dichotomized items are as follows:

Table OA22: At max at first post-treatment

	With Covariates			
	Effect	SE	t.stat	p
law_appointment (P*)	-0.045	0.026	-1.757	0.079
law_private_insur(P*)	0.020	0.020	1.014	0.311
law_counsel (P*)	-0.029	0.024	-1.210	0.227
law_public_insur (P)	-0.023	0.026	-0.898	0.370
ballot_abortion (P*)	-0.031	0.025	-1.234	0.217
abortion_something_wrong (S*)	0.017	0.014	1.161	0.246
abortion_badly (S*)	0.027	0.010	2.660	0.008
abortion_consider (S)	-0.014	0.023	-0.604	0.546
abortion_birthcontrol (S*)	0.020	0.017	1.172	0.242
abortion_nothing_wrong (S)	0.004	0.025	0.154	0.877
act_volunteer_clinic (A)	0.006	0.014	0.414	0.679
act_accompany (A)	0.069	0.028	2.437	0.015
act_protest_clinic (A*)	0.006	0.008	0.772	0.441
act_congress_support (A)	0.063	0.019	3.289	0.001
act_congress_opposition (A*)	0.022	0.016	1.360	0.174

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 are coded to 1 if a participant agreed at all with the policy/statement and 0 otherwise (indicating stated indifference or opposition). For the action-taking items, we examine cases where the respondent was either extremely or very likely. The results on the individual items dichotimized in this manner are as follows:

Table OA23: Agree at all at first post-treatment

	With Covariates			
	Effect	SE	t.stat	p
law_appointment (P*)	-0.066	0.031	-2.147	0.032
law_private_insur(P*)	-0.001	0.024	-0.060	0.952
law_counsel (P*)	-0.059	0.028	-2.114	0.035
law_public_insur (P)	0.021	0.028	0.766	0.444
ballot_abortion (P*)	-0.024	0.027	-0.897	0.370
abortion_something_wrong (S*)	0.040	0.023	1.765	0.078
abortion_badly (S*)	0.011	0.018	0.618	0.537
abortion_consider (S)	0.019	0.029	0.650	0.516
abortion_birthcontrol (S*)	-0.026	0.027	-0.966	0.334
abortion_nothing_wrong (S)	0.006	0.030	0.193	0.847
act_volunteer_clinic (A)	0.015	0.018	0.828	0.408
act_accompany (A)	0.077	0.031	2.473	0.014
act_protest_clinic (A*)	0.005	0.010	0.529	0.597
act_congress_support (A)	0.090	0.024	3.753	0.000
act_congress_opposition (A*)	0.025	0.021	1.207	0.228

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 [6] 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 [5] and weight on gender and age (we do not have vote history and the universe is overwhelmingly white; we have no other covariates on the starting universe).

Below are results with and without these weights, showing that the estimated SATEs and PATEs are similar.

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

Table OA24: ATE Effects with Weights

	Effect	SE	t.stat	p
Policy				
Unweighted - 1 Wk	0.061	0.032	1.933	0.054
Weighted - 1 Wk	0.066	0.030	2.189	0.029
Unweighted - 1 Mo	0.040	0.033	1.216	0.224
Weighted - 1 Mo	0.045	0.033	1.382	0.167
Unweighted - 3 Mo	0.032	0.036	0.878	0.380
Weighted - 3 Mo	0.035	0.037	0.955	0.340
Stigma				
Unweighted - 1 Wk	0.033	0.038	0.847	0.397
Weighted - 1 Wk	0.047	0.037	1.266	0.206
Unweighted - 1 Mo	0.022	0.037	0.595	0.552
Weighted - 1 Mo	0.031	0.038	0.801	0.423
Unweighted - 3 Mo	0.045	0.038	1.194	0.233
Weighted - 3 Mo	0.059	0.039	1.511	0.131
Action				
Unweighted - 1 Wk	0.186	0.051	3.663	0.000
Weighted - 1 Wk	0.195	0.053	3.714	0.000
Unweighted - 1 Mo	0.105	0.052	2.034	0.042
Weighted - 1 Mo	0.069	0.055	1.258	0.209
Unweighted - 3 Mo	0.080	0.054	1.501	0.134
Weighted - 3 Mo	0.093	0.057	1.633	0.103

Comparison with Broockman, Kalla, and Sekhon (2017)

This study and [2] have very different populations. For example, this study is nearly 100% white while [2] is around 46% white. This could potentially explain the differences in the findings between the two studies. To address this concern, we re-analyzed [2] by re-weighting it to match the demographics of the current study. In particular, we used entropy balancing [5] and race, gender, age, partisanship, and ideology. On the All DVs t1 outcome, [2] report (see their Table OA4, Model 1, Row 1) an effect of -0.01 (SE = 0.04). In our re-analysis with re-weighting, we estimate an effect of 0.003 (SE = 0.05). After re-weighting [2] to match the demographics of the current study, we continue to find a null effect in [2].[1]

[1] We thank an anonymous reviewer for this helpful suggestion.

References

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

[2] 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.

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

[4] 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.

[5] 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.

[6] 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.

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