CHESS: Continuous Heterogeneity Estimation for Self-Selection into Treatment Using Willingness-to-Pay

Karthik Srinivasan[*]
University of Michigan
School of Information

First Draft: September 8, 2025
This Draft: September 10, 2025

Abstract

In many settings, individuals choose whether to receive a treatment, but experimental designs often ignore this self-selection. This paper introduces a method—Continuous Heterogeneity Estimation for Self-Selection (CHESS)—that uses incentive-compatible willingness-to-pay elicitation to estimate treatment effect heterogeneity as a function of selection propensity. Unlike prior approaches that incorporate binary participant choice into the experimental design (e.g., PICA), CHESS enables estimation of continuous, non-linear heterogeneity and avoids sensitivity to researcher-supplied alternatives to treatment. In a preregistered online experiment with 812 respondents, I use CHESS to study gun control issue advertisements. An anti-gun control ad reduces support for gun control by 0.27 standard deviations, while a pro-gun control ad increases issue priority by 0.26 standard deviations among the respondents most likely to view it. This proof-of-concept demonstrates that CHESS provides a flexible framework for analyzing heterogeneity when self-selection is central to treatment assignment.

[*] You can contact me at ksrini@umich.edu. This paper is a direct result of conversations with my coauthors Scott Behmer and Rafael Jiménez-Durán. I would like to thank Nicholas Valentino, Katie Nissen and participants at the Interdisciplinary Workshop on Comparative Politics and the Behavioral and Experimental Economics Lab for valuable feedback. Additionally, I would like to thank my family and friends including, but not limited to, Amma, Appa, Arjun, Anand, Sam, Anna, Gabi, Alex, Jaclyn, Axel, Walter and Mochi.

Campaigns spend millions of dollars testing advertisements through randomized controlled trials. During the 2024 presidential campaign, one political consultant tested "4,000 or 5,000 Harris ads" explaining "for any given ad, you take a thousand people, split them into treatment and control, and then compare outcomes" (Klein, 2025). This design captures the causal effect of exposure under random assignment. Yet, advertisements can only persuade the people who choose to watch them. The relevant question to a campaign is not the average treatment effect, but the treatment effect among those who would watch the ad. Standard advertising RCTs, by forcing exposure, abstract away from self-selection into treatment.

This problem extends far beyond campaign advertising. Medical treatments can only affect patients who consent to treatment. Social services can only reach participants who sign up. Across domains, interventions typically require individuals to self-select into treatment. Researchers and policymakers are often interested less in the average treatment effect than in the effect among those most likely to be treated.

Existing experimental designs provide one way to address this problem. The Preference-Incorporating Choice and Assignment (PICA) design, also known in medical research as the patient preference trial (PPT), incorporates participant choice directly into the experiment (de Benedictis-Kessner et al., 2019). In PICA, some participants are assigned to a forced-exposure arm, where treatment is randomized as in a conventional RCT. Others are assigned to a choice arm, where they can either select into treatment or opt for an outside option. In this way, PICA identifies treatment effects separately for those who would and would not select treatment.

However, PICA has two important limitations. First, the design treats a continuous variable, the propensity to self-select into treatment, as a binary variable. This restricts the researcher to estimating only two treatment effects: one for those who choose the treatment, and one for those who choose the outside option. Second, the conclusions drawn from PICA can shift dramatically depending on the desirability of the researcher-supplied outside option, in some cases even reversing who appears to benefit from treatment.

This paper introduces an alternative experimental design—Continuous Heterogeneity Estimation for Self-Selection (CHESS)—that resolves these limitations. The idea behind CHESS is to directly estimate a measure of participants' willingness to self-select into treatment. To accomplish this, CHESS uses an incentive-compatible elicitation procedure from economics, the Becker–DeGroot–Marschak (BDM) method, to measure participants' willingness to pay (WTP) for treatment (Becker et al., 1964). Intuitively, this procedure asks: What is the minimum amount you would need to be paid in order to agree to treatment? Because the elicitation is incentive compatible, participants have reason to answer truthfully. This dollar value provides a natural, continuous measure of each individual's propensity to self-select into treatment. By combining this measure with random assignment, CHESS allows researchers to estimate treatment effect heterogeneity continuously, capture non-linear patterns, and avoid dependence on a researcher-supplied outside option.

I illustrate CHESS with a preregistered online experiment (N=812) studying pro- and anti-gun control advertisements. The results demonstrate the value of this experimental design. While the average treatment effect suggests that a pro-gun control ad has little effect overall, CHESS shows that it increases the issue priority of gun control among the respondents most likely to view it. In contrast, an anti-gun control ad consistently decreases support for gun control across the selection spectrum. These results show how CHESS can provide insights that conventional designs miss through a more granular analysis of heterogeneity by willingness-to-select-into-treatment.

1 Overview of CHESS Experimental Design

Suppose we want to estimate heterogeneity by willingness-to-select-into-treatment. If this quantity were observable, the task would be straightforward: after randomizing participants into treatment and control, we could simply estimate

$$\mathbb{E}[Y_1 - Y_0 \mid W = w],$$

for each level w of willingness-to-select-into-treatment (for example, high, medium, or low). This logic is identical to how researchers estimate heterogeneity along other observable dimensions such as gender, race, or education. The difficulty is that willingness-to-select-into-treatment is typically unobserved.

The CHESS design addresses this problem by reframing the question from "How likely would you be to select into treatment?" to "What is the minimum amount of money you would need to be paid in order to participate in treatment?" Directly asking participants this question risks mismeasurement, as participants might strategically under- or over-report.[1] To mitigate this problem, CHESS elicits willingness-to-pay (WTP) using an incentive-compatible elic-

[1] For example, participants may believe that overstating WTP increases payment, or understating it increases the chance of selection.

itation mechanism, the Becker–DeGroot–Marschak (BDM) procedure (Becker et al., 1964; Myerson, 1979). Because the mechanism is incentive compatible, participants have an incentive to report their WTP truthfully. The elicited value then provides a continuous measure of each individual's propensity to self-select into treatment.

The CHESS design proceeds in three steps:

1. Elicit an answer to the question "How much would I need to pay you in order for you to participate in treatment?" using an incentive-compatible mechanism.
2. Randomize participants into treatment and control, as in a standard experiment.
3. Estimate treatment effect heterogeneity by grouping participants according to their elicited values of willingness-to-select-into-treatment.

This approach yields a continuous measure of selection propensity and enables researchers to recover patterns of treatment effect heterogeneity that binary choice designs cannot.

2 Literature

This paper contributes to the literature on experimental designs that incorporate self-selection into treatment. Rücker (1989) proposes the idea of an experimental design with forced-exposure and selected-exposure arms. This type of design has take-up in the medical literature, where it is referred to as a patient-preference trial (Torgerson and Sibbald, 1998; King et al., 2005). Gaines and Kuklinski (2011) propose an experimental design that uses forced- and selective-exposure arms to study the effect of attack ads. Arceneaux et al. (2012) use separate forced-exposure, selective-exposure and stated-preference experiments to study the effects of counter-attitudinal news. Knox et al. (2019) proposes a variant of the forced and selective exposure design where stated preferences are initially elicited from all participants before randomization to study the effects of partisan media exposure. de Benedictis-Kessner et al. (2019) study the persuasive effect of partisan media by eliciting stated and revealed preferences, incorporating forced and selective-exposure arms to allow for the quantification of measurement error.[2]

[2] The broader problem of estimating heterogeneous treatment effects in the context of selection into treatment has long been formally studied in the econometrics literature (Heckman and Smith, 1995; Angrist et al., 1996).

This paper leverages the willingness-to-pay elicitation method proposed in Becker et al. (1964) with an implementation from Berkouwer and Dean (2022).

This paper also relates to a large literature on the persuasive effects of political advertising. Specifically, this paper contributes to the strand of the literature focused on experimental evaluations of issue advertising. Kalla and Broockman (2022) study issue ads regarding immigration and LGBTQ inclusion, and find that ads are remembered but largely unpersuasive. Junk and Rasmussen (2024) study an advertising campaign about privacy risk, and find that the ad does not sway public opinion or make the issue more salient, but does change intended consumer behavior. The prior literature theorizes that issue advertising may operate via increasing the salience of an issue or changing public attitudes, which motivates the two outcome variables collected in this study (Hall and Reynolds, 2012).

3 Statistical Framework Connecting PICA and CHESS

In order to more precisely compare the CHESS and PICA designs, I now formalize the estimation problem using a potential outcomes framework (Neyman, 1923; Rubin, 1974). Let $X_i \in \{0, 1\}$ indicate whether individual i receives treatment, let Y_i denote individual i's observed outcome, and let $C_i \in \{0, 1\}$ denote whether individual i chooses to participate in the intervention if given a choice. Let $Y_i(1)$ and $Y_i(0)$ denote individual i's potential outcomes under treatment and control, respectively.

In a standard PICA design, participants are randomly assigned to one of two arms:

1. A *choice arm*, where participants self-select into the intervention (C_i is observed and determines X_i);
2. A *forced-assignment arm*, where participants are assigned to treatment or control regardless of preference (X_i is randomly assigned).

This design allows us to identify the following treatment effects:

1. $E[Y_i(1) - Y_i(0)]$, the average treatment effect
2. $E[Y_i(1) - Y_i(0) \mid C_i = 1]$, the treatment effect among participants who would opt into treatment if given a choice

3. $E[Y_i(1) - Y_i(0) \mid C_i = 0]$, the treatment effect among participants who would opt into control if given a choice

Let z denote a participant's willingness-to-select-into-treatment, which we can think of as their anticipated utility from participation.

In the choice arm, the researcher offers participants the choice to either opt into treatment or to select an outside option. For now, assume that all participants value the outside option at a fixed level, d. Then,

$$C_i = \mathbb{I}[z_i > d]$$

That is, participants opt into treatment if they like treatment more than the outside option. Implicitly, the provided alternative divides participants into high- and low-willingness-to-select-into-treatment groups. Then, the PICA heterogeneous treatment effects correspond to

$$\begin{gathered} E[Y_i(1) - Y_i(0) \mid z > d] \\ E[Y_i(1) - Y_i(0) \mid z \leq d], \end{gathered}$$

where the value of d depends on how desirable the researcher-supplied outside option is to participants.

Suppose that the elicitation procedure provides a valid measure of participants' willingness-to-select-into-treatment. That is, assume that the answer to "How much would I need to pay you to participate in treatment?" perfectly correlates with willingness-to-select-into-treatment z_i. Then, we can compute

$$E[Y_i(1) - Y_i(0) \mid z \in (a, b)]$$

for any interval (a, b) with empirical support in the data. In particular, we can recover the PICA estimates, as well as any other intervals of interest.

In practice, we may want to provide a non-parametric estimate for how the treatment effect varies with willingness-to-select-into-treatment. To do this, we can define cutpoints along z that divide the data into K evenly sized bins, and then can compute K within-bin treatment effects. In this way, CHESS allows researchers to estimate how treatment effects vary across the full distribution of selection propensity, rather than being restricted to a binary analysis as in PICA.

4 Comparing CHESS and PICA Designs

At a high level, CHESS and PICA estimate heterogeneous treatment effects by willingness-to-select-into-treatment. The key difference is that CHESS treats willingness-to-select-into-treatment as continuous, while PICA treats it as binary. Intuitively, this means that CHESS enables a more fine-grained analysis of treatment effect heterogeneity, but this comes at the cost of a more complicated elicitation procedure that requires more observations to achieve the same degree of precision.

Because PICA elicits willingness-to-select-into-treatment as a binary variable, it restricts the researcher to estimating only two heterogeneous treatment effects: one for those who opt into treatment in the choice arm, and one for those who opt out. In contrast, CHESS allows the researcher to estimate a heterogeneous treatment effect for any interval along the willingness-to-select-into-treatment continuum. Notably, this allows CHESS to identify non-linear heterogeneous treatment effects, which PICA cannot do.

A second, tightly related issue is that the conclusions of a PICA analysis may depend on the researcher-supplied alternative to treatment in the choice arm. Figure 13 illustrates this problem with a hypothetical example. In this example, heterogeneous treatment effects are non-linear and largest for participants with a moderate willingness-to-select-into-treatment. The second and third panels of Figure 13 simulate the heterogeneous treatment effects estimated by PICA when the researcher-supplied alternative to treatment is relatively more or less desirable.[3] In this example, changing the desirability of the alternative reverses the conclusion of the PICA analysis. In contrast, CHESS recovers a continuous treatment effect curve, which prevents this issue.

The advantages of CHESS come with two main costs. First, eliciting willingness-to-pay involves an additional step that may confuse respondents.[4] This confusion can reduce response quality and require extra comprehension checks, which lengthen the survey. CHESS also complicates compensation, since participants receive different payments depending on their responses. In contrast, PICA offers all participants a straightforward and interpretable choice.

[3] In panel two, the outside option is more desirable, so those with moderate willingness-to-select-into-treatment opt out, leading the researcher to conclude that the treatment is less effective for those who are likely to select into treatment. In panel three, the outside option is less desirable, so those with moderate willingness-to-select-into-treatment opt in, leading to the opposite conclusion.

[4] There are a few common willingness-to-pay elicitation procedures including standard BDM, multiple price lists, and binary search, each of which have costs and benefits. The stability and accuracy of willingness-to-pay elicitation mechanisms is a subject of discussion.

Second, to achieve comparable precision, CHESS requires more data. Specifically, in Appendix A, I show that estimating K heterogeneous treatment effects with CHESS requires $\frac{K}{2}$ times the sample size of PICA to achieve comparable precision.[5]

5 Application: Evaluating Gun Control Issue Ads

I conducted a preregistered online survey experiment on Prolific in September 2025.[6] A total of 883 respondents were recruited, restricted to U.S. residents aged 18-65.[7]

Following the CHESS design, I elicit participants' willingness-to-pay to view a one-minute ad. Using a binary search procedure, I ask participants "Would you be willing to watch a 1 minute long advertisement for [price]?" searching the interval from 10 cents to 2 dollars in 10 cent increments. Participants were recruited into treatment if their willingness-to-pay fell below the randomly chosen price. In practice, nearly all participants were compared against a $1.00 price, which resulted in 812 of the 883 respondents being recruited into treatment.[8]

Treated respondents were randomly assigned to one of three conditions, which varied the ad that participants viewed. In the pro-gun control treatment condition, participants viewed Back-To-School Essentials, an ad created by Sandy Hook Promise. In the anti-gun control treatment, participants viewed Freedom's Safest Place — Real Empowerment, an ad created by the National Rifle Association. In the placebo condition, participants viewed a PSA about texting while driving produced by the news station WJBF. Participants completed the outcome measures immediately after treatment exposure.

There are two outcome variables: issue attitude and issue priority. Issue attitude is the respondent's support for gun control, as measured by a 7-point Likert scale question "In your opinion, should it be harder or easier to get a gun in this country?" with higher values indicating greater support for gun control. Issue priority is defined by the rank of "gun laws" in response to the question "Please rank the following issues from 1 (most important)

[5] One way to relax the demand that the sample size of CHESS must linearly increase with the number of bins is to make a functional form assumption regarding the shape of heterogeneous treatment effects (e.g. a second degree polynomial). In this case, all data can be applied to recover the parametric estimates.

[6] Preregistration is available at https://aspredicted.org/fknm-9qxz.pdf

[7] I also place a few additional technical restrictions on Prolific recruitment: I restrict to prolific workers with a 99%+ approval rate on previous studies who have not participated in any of this study's pilots.

[8] The randomization procedure is $1.00 with 99% probability, and $2.00 with 1% probability. I do this so that the description that the price is randomly chosen as well as the claim that the bonus may be up to $2.00 is truthful.

to 5 (least important) when it comes to deciding your vote. The ranking should reflect how much the issue matters to you, no matter what position you personally take on it." with the options (a) The Environment/Climate Change, (b) Immigration Policy, (c) Gun Laws, (d) Abortion Policy, and (e) The Economy.[9]

I preregister three hypotheses:

- H1: Relative to the placebo ad, the pro and anti-gun control ads will raise the issue priority of gun control.
- H2: Relative to the placebo ad, the pro-gun control ad will increase support for gun control.
- H3: Relative to the placebo ad, the anti-gun control ad will decrease support for gun control.

Figure 1 and Table 1 report the main treatment effects. Overall, while all point estimates are directionally consistent with each of these hypotheses, I do not find statistically significant treatment effects on issue priority for either ad. The estimated effect of the pro-gun control ad on support for gun control is close to zero (0.16 units, 95% CI: [-0.08,0.40]). The anti-gun control ad reduces support for gun control by 0.37 units (95% CI: [-0.13,-0.61]). This is a modest, but meaningful treatment effect: it represents 27% of the baseline difference between liberal and conservative attitudes on gun control (mean: 6.27 vs. 4.94), or about one quarter of a standard deviation (SD = 1.37).

The primary purpose of this experiment is to demonstrate how CHESS enables estimating heterogeneous treatment effects by self-selection into treatment. I use the elicited WTP measure, which is the minimum price that I need to pay a participant in order for them to watch an ad, as a proxy for willingness-to-select-into-treatment. In Figure 2, Table 2, and Table 3, I report heterogeneous treatment effects splitting propensity-to-select-into-treatment into three bins (high, medium and low propensity). While the treatment effects on support for gun control are largely consistent across selection propensity, the pro-gun control ad increases issue priority by 0.33 ranks (95% CI: [0.01-0.66]) only for high propensity respondents. This is a modest treatment effect, representing 28% of a standard deviation in the rank variable (SD = 1.19). This finding exemplifies the usefulness of CHESS: overall treatment effects suggested that the pro-gun control ad was ineffective, but the ad turns out

[9] For presentation in figures and tables, I reverse the scale, so that higher values indicate greater importance.

to have a meaningful effect on issue priority for the respondents who are most likely to watch it according to this measure of self-selection into treatment.

Beyond these main results, I preregister a variety of secondary analyses. Figures 3 and 4 document two alternative approaches to using the self-selection measure to analyze heterogeneity. Figure 3 is a coarse 2-bin analysis, comparable to the PICA framework, while Figure 4 is a continuous approach interacting treatment with a quadratic in willingness-to-pay. While this study is not particularly well-powered for estimating fine-grained heterogeneity, both figures suggest similar overall takeaways: the anti-gun control ad decreases support for gun control, while the pro-gun control ad does not increase support to a statistically significant degree. However, the pro-gun control ad increases the issue priority of gun control for high-propensity respondents.

Figure 5 documents heterogeneity in treatment effects by political ideology. There are not meaningful differences across ideology in the issue priority treatment effects, but the anti-gun control ad is more effective at moving support for gun control among conservatives. Figure 6 interacts political party with with (binary) propensity to select into treatment. This cut more clearly shows that the pro gun control ad moves both attitudes towards gun control and issue priority for high-propensity liberals, but does not influence conservatives or low-propensity liberals. This analysis again highlights how splitting by propensity to self-select into treatment can be valuable: the pro-gun control ad appears a lot more effective among the respondents who are most likely to watch it according to this measure of self-selection.

While not the primary purpose of this experiment, I also provide (preregistered) heterogeneity by political engagement (Figure 7), race (Figure 8), gender (Figure 9), and educational attainment (Figure 10).

Figure 11 presents a simple analysis of respondents free-responses to the question "Did you find the ad persuasive? Why or why not?" Responses were coded as persuasive or not persuasive using a large language model (GPT-4), with coding validated against human coding for a random subset. Interestingly, respondents perceive the pro-gun control ad as much more persuasive than the anti-gun control ad, though the causal treatment effects show that this is not the case in practice.

Lastly, I show that the main results are robust to dropping respondents who fail an additional attention check focused on the content of the ads (Figure 12).

Overall, these results demonstrate how the CHESS design enables a flexible analysis of heterogeneity by the propensity to self-select into treatment. Moreover, this example demonstrates how this kind of analysis can yield valuable insights that would not be recovered by a standard RCT, and which lead to different conclusions regarding treatment efficacy.

6 Conclusion

I propose a new experimental design, Continuous Heterogeneity Estimation for Self-Selection (CHESS). This research design highlights a connection between an old literature in economics on eliciting willingness-to-pay and the more recent work in political science and medical science focused on leveraging participant choice to uncover heterogeneity by self-selection (e.g., PICA, PPT).

While PICA restricts researchers to estimating binary heterogeneous treatment effects, CHESS treats self-selection propensity as a continuous variable. This enables CHESS to overcome two key disadvantages of PICA designs: (1) PICA cannot estimate non-linear treatment effects and (2) the conclusions of PICA designs are sensitive to the researcher-supplied alternative to treatment. However, these advantages come at the cost of requiring a larger sample and increasing experimental complexity.

I demonstrate the design with an application to political advertising. CHESS reveals that a pro-gun control ad modestly increases the issue priority of gun control among respondents who are most likely to select into treatment, while an anti-gun control ad has consistent negative effects on attitudes towards gun control across the selection spectrum.

When researchers can support the larger samples required and when participant self-selection is an important concern, CHESS can provide more granular insights into treatment effect heterogeneity.

References

Angrist, J. D., G. W. Imbens, and D. B. Rubin (1996). Identification of causal effects using instrumental variables. *Journal of the American statistical Association 91*(434), 444–455.

Arceneaux, K., M. Johnson, and C. Murphy (2012). Polarized political communication, oppositional media hostility, and selective exposure. *The Journal of Politics 74*(1), 174–186.

Becker, G. M., M. H. DeGroot, and J. Marschak (1964). Measuring utility by a single-response sequential method. *Behavioral science 9*(3), 226–232.

Berkouwer, S. B. and J. T. Dean (2022). Credit, attention, and externalities in the adoption of energy efficient technologies by low-income households. *American Economic Review 112*(10), 3291–3330.

de Benedictis-Kessner, J., M. A. Baum, A. J. Berinsky, and T. Yamamoto (2019). Persuading the enemy: Estimating the persuasive effects of partisan media with the preference-incorporating choice and assignment design. *American Political Science Review 113*(4), 902–916.

Gaines, B. J. and J. H. Kuklinski (2011). Experimental estimation of heterogeneous treatment effects related to self-selection. *American Journal of Political Science 55*(3), 724–736.

Hall, R. L. and M. E. Reynolds (2012). Targeted issue advertising and legislative strategy: The inside ends of outside lobbying. *The Journal of Politics 74*(3), 888–902.

Heckman, J. J. and J. A. Smith (1995). Assessing the case for social experiments. *Journal of economic perspectives 9*(2), 85–110.

Junk, W. M. and A. Rasmussen (2024). Are citizens responsive to interest groups? a field experiment on lobbying and intended citizen behaviour. *West european politics 47*(7), 1643–1669.

Kalla, J. L. and D. E. Broockman (2022). “outside lobbying” over the airwaves: A randomized field experiment on televised issue ads. *American Political Science Review 116*(3), 1126–1132.

King, M., I. Nazareth, F. Lampe, P. Bower, M. Chandler, M. Morou, B. Sibbald, and R. Lai (2005). Impact of participant and physician intervention preferences on randomized trials: a systematic review. *Jama 293*(9), 1089–1099.

Klein, E. (2025, March). Democrats need to face why trump won. The Ezra Klein Show, The New York Times. Transcript of interview with David Shor.

Knox, D., T. Yamamoto, M. A. Baum, and A. J. Berinsky (2019). Design, identification, and sensitivity analysis for patient preference trials. *Journal of the American Statistical Association 114*(528), 1532–1546.

Myerson, R. B. (1979). Incentive compatibility and the bargaining problem. *Econometrica: journal of the Econometric Society*, 61–73.

Neyman, J. (1923). On the application of probability theory to agricultural experiments. essay on principles. *Ann. Agricultural Sciences*, 1–51.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology 66*(5), 688.

Rücker, G. (1989). A two-stage trial design for testing treatment, self-selection and treatment preference effects. *Statistics in medicine 8*(4), 477–485.

Torgerson, D. J. and B. Sibbald (1998). Understanding controlled trials. what is a patient preference trial? *BMJ: British Medical Journal 316*(7128), 360.

A Precision of CHESS/PICA Estimators

In this section, I'll show that, under some simplifying assumptions, a researcher who wants to devide the willingness-to-select-into-treatment spectrum into K bins must collect $\frac{K}{2}$ times as many observations to estimate CHESS treatment effects with the same degree of precision as PICA.

A.1 Standard Error of Treatment Effect Estimator

First, I will derive the standard error of a treatment effect estimator. Let $X \in \{0, 1\}$ indicate the treatment condition, and let Y be the outcome variable. Suppose we have N observations, randomized evenly into treatment and control. The treatment effect estimate is given by

$$\tau = \mathbb{E}[Y|X = 1] - \mathbb{E}[Y|X = 0]$$

with estimator

$$\hat{\tau} = \frac{2}{N}\sum_{i=1}^{N} y_i \mathbb{I}[X = 1] - \frac{2}{N}\sum_{i=1}^{N} y_i \mathbb{I}[X = 0].$$

Then,

$$\text{Var}(\hat{\tau}) = \text{Var}(\bar{Y}_1 - \bar{Y}_0) = \text{Var}(\bar{Y}_1) + \text{Var}(\bar{Y}_0).$$

Define

$$\sigma_0^2 = \text{Var}(Y|X = 0) \qquad \sigma_1^2 = \text{Var}(Y|X = 1).$$

Then,

$$\text{Var}(\bar{Y}_1) = \frac{2\sigma_1^2}{N} \qquad \text{Var}(\bar{Y}_0) = \frac{2\sigma_0^2}{N}$$

Assuming homoskedasticity of errors,

$$\sigma^2 = \sigma_0^2 = \sigma_1^2$$

so

$$\text{Var}(\hat{\tau}) = \text{Var}(\bar{Y}_1) + \text{Var}(\bar{Y}_0) = \frac{2\sigma_1^2}{N} + \frac{2\sigma_0^2}{N} = \frac{4\sigma^2}{N}$$

and

$$\mathrm{SE}(\hat{\tau}) = 2\sigma\sqrt{\frac{1}{N}}.$$

A.2 K Heterogeneous Treatment Effect Bins

Now, I will derive the standard error for heterogeneous treatment effects when we divide N observations into K equal sized bins, representing K equal cuts along some dimension of heterogeneity Z. In each bin, there are $\frac{N}{K}$ observations. As a simplifying assumption, I impose homoskedasticity of errors within each of these bins. Formally, $\forall k \in K$,

$$\sigma^2 = \mathrm{Var}(Y \mid X = 0, Z \in \text{bin } k) = \mathrm{Var}(Y \mid X = 0, Z \in \text{bin } k)$$

Then,

$$\mathrm{Var}(\hat{\tau}_k) = \frac{4K\sigma^2}{N}$$

and

$$\mathrm{SE}(\hat{\tau}_k) = 2\sigma\sqrt{\frac{K}{N}}$$

If the PICA outside option happens to split the sample exactly in half (a best case scenario, ensuring the most precise estimates), then PICA is equivalent to $K = 2$, so

$$\mathrm{SE}(\hat{\tau}_{\text{PICA}}) = 2\sigma\sqrt{\frac{2}{N}}$$

In the CHESS case, the researcher can divide the willingness-to-select spectrum into K equally-sized bins, so K is a parameter to be set by the researcher, and the efficiency is given by

$$\mathrm{SE}(\hat{\tau}_{\text{CHESS}}) = 2\sigma\sqrt{\frac{K}{N}}$$

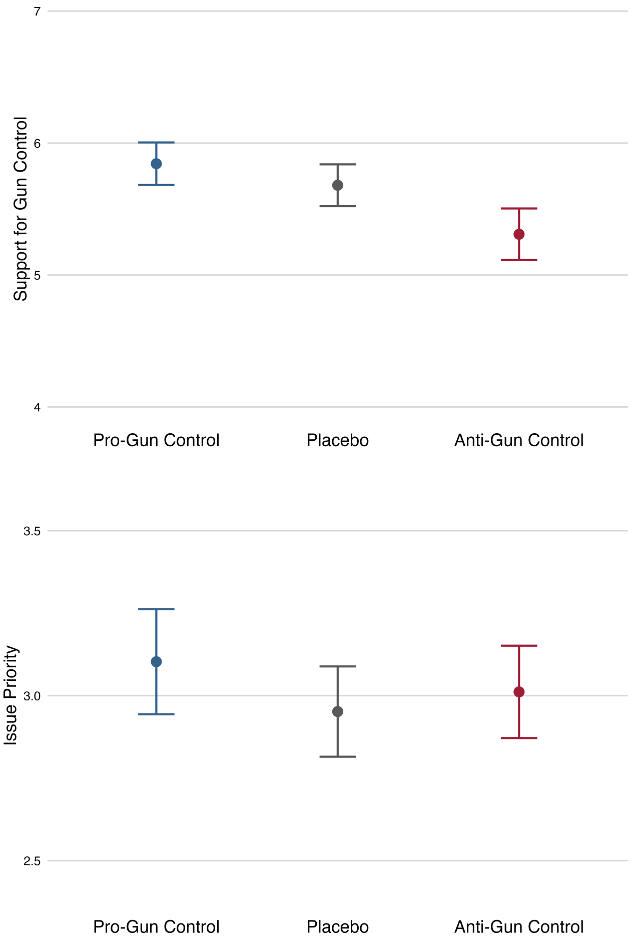
Then, suppose we have N_{PICA} observations for a PICA design and we select K groups for CHESS. I solve for the number of observations N^*_{CHESS} required to estimate treatment effects with equal precision to PICA:

$$\begin{aligned}
\mathrm{SE}(\hat{\tau}_{\text{PICA}}) &= \mathrm{SE}(\hat{\tau}_{\text{CHESS}}) \\
2\sigma\sqrt{\frac{2}{N_{\text{PICA}}}} &= 2\sigma\sqrt{\frac{K}{N_{\text{CHESS}}}} \\
\frac{N_{\text{CHESS}}}{N_{\text{PICA}}} &= \frac{K}{2} \\
N^{*}_{\text{CHESS}} &= \frac{K}{2} N_{\text{PICA}}
\end{aligned}$$

So, we need $\frac{K}{2}$ times as much data for CHESS to achieve the same level of precision as PICA.

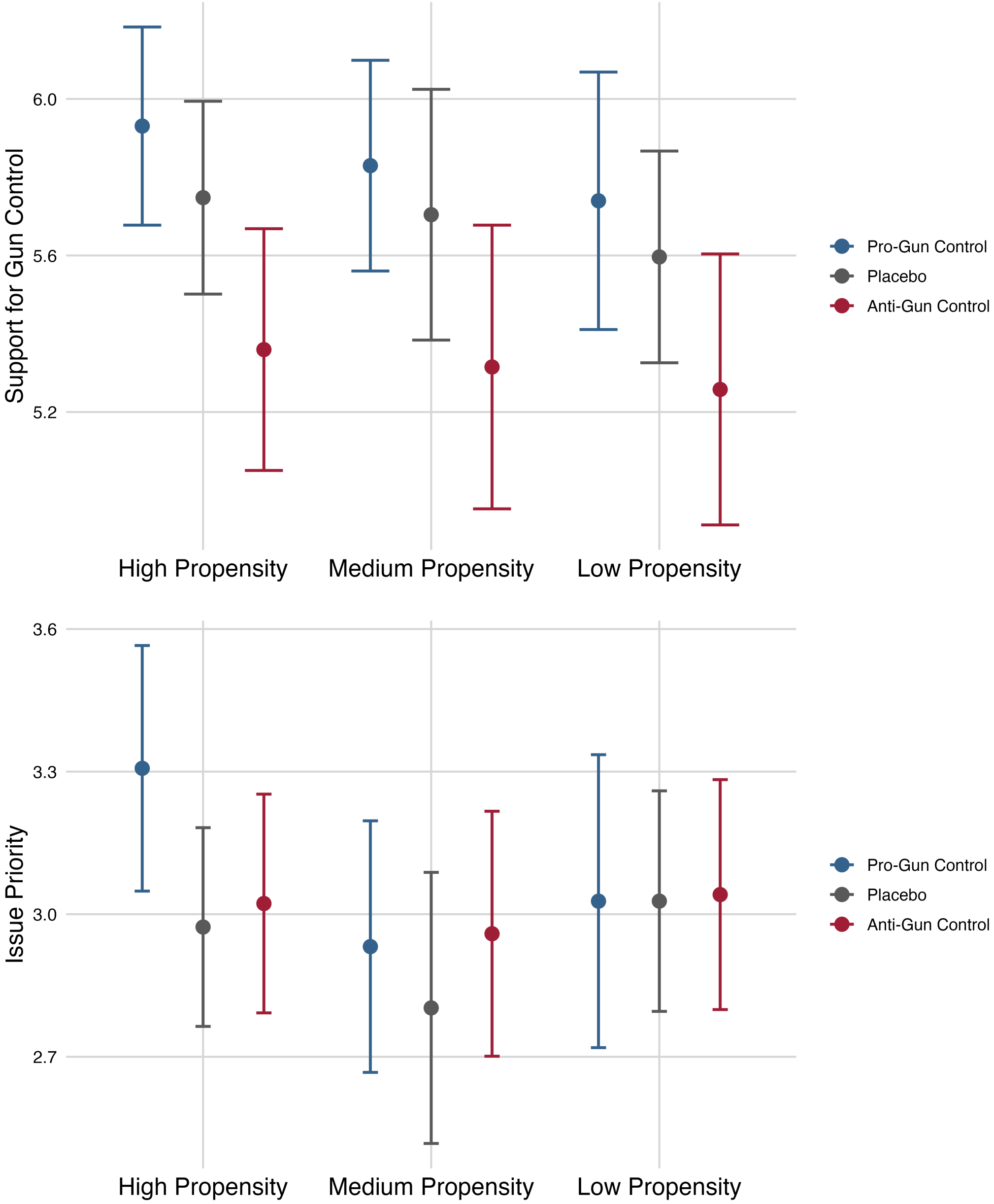
B Figures

Figure 1: Average Outcomes by Treatment Arm



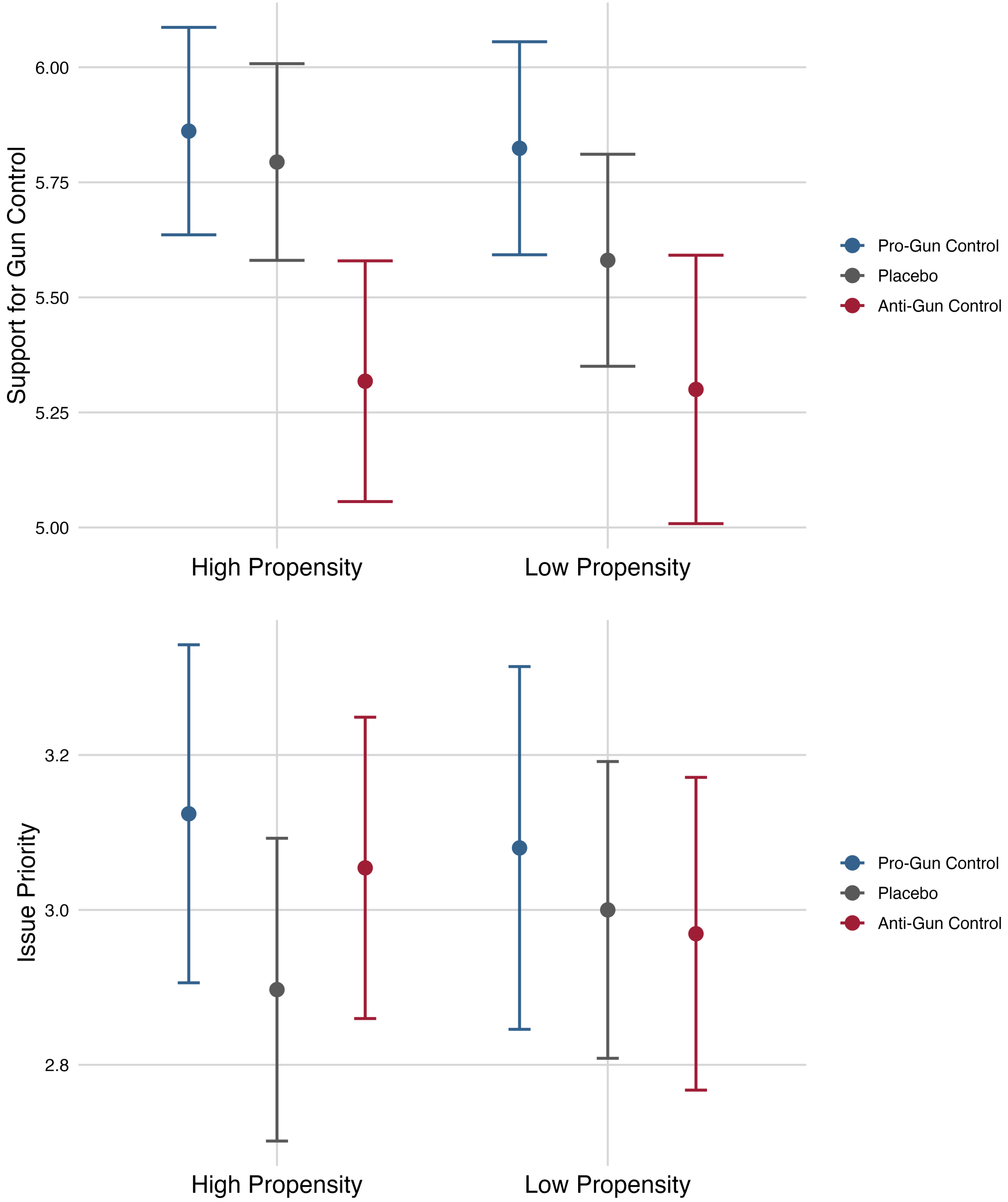
Note: The first panel plots the support for gun control, which is the response to “In your opinion, should it be harder or easier to get a gun in this country?” on a 7-point Likert scale, with 7 representing “Much harder to get a gun.” The second panel plots the relative issue priority of gun control, which is the ranking of “gun laws” in response to “Please rank the following issues from 1 (most important) to 5 (least important) when it comes to deciding your vote. The ranking should reflect how much the issue matters to you, no matter what position you personally take on it.” with the options (a) The Environment/Climate Change, (b) Immigration Policy, (c) Gun Laws, (d) Abortion Policy, (e) The Economy. For graphing, I reverse the rank so that higher values represent more importance. Points represent means, and error bars represent 95% confidence intervals.

Figure 2: Heterogeneity by Self-Selection into Treatment



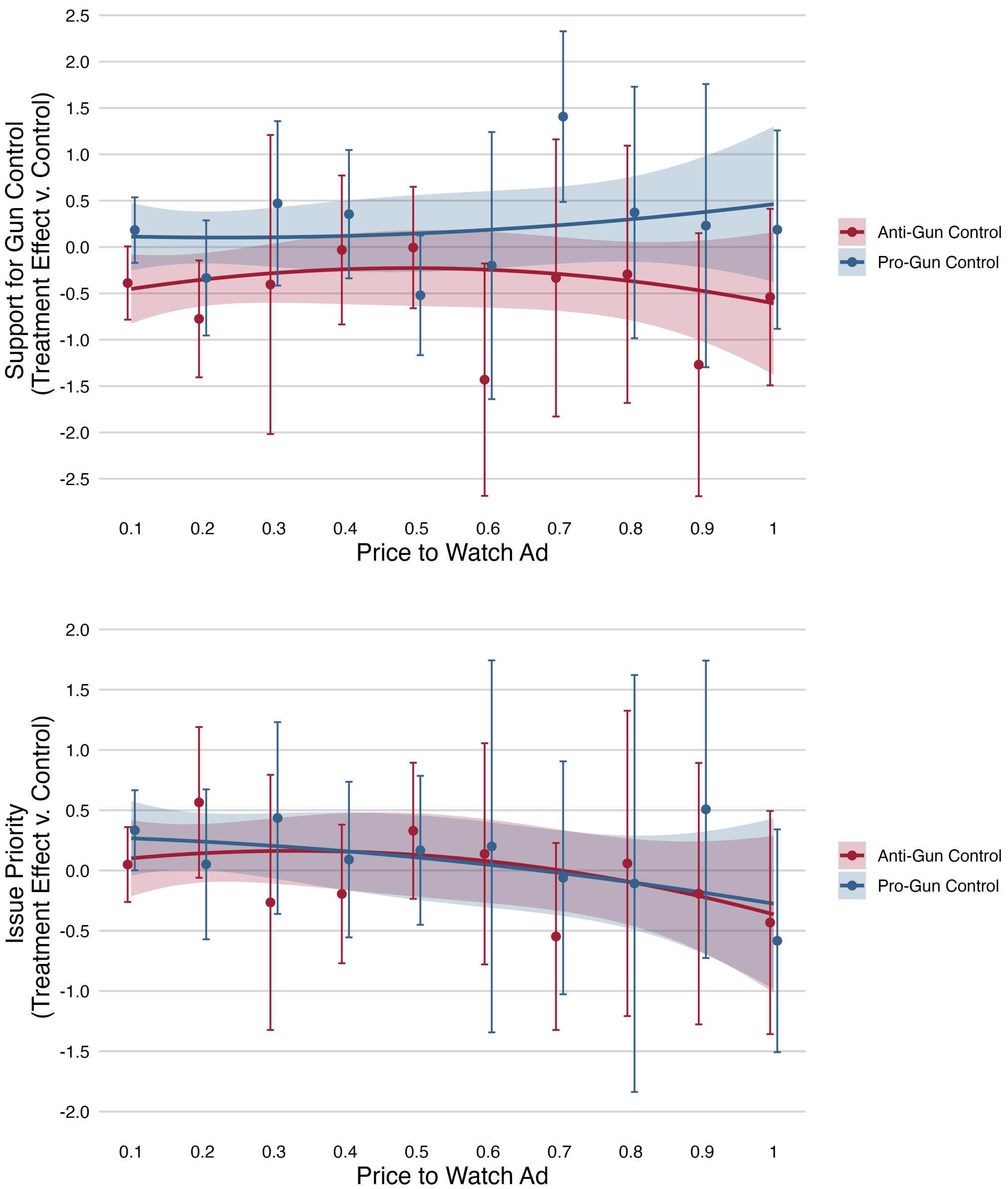
Note: This figure represents the main methodological innovation of this paper: it plots heterogeneity by a measure of self-selection into treatment. Points represent means, and error bars represent 95% confidence intervals.

Figure 3: Heterogeneity by Self-Selection into Treatment (2 Bins)



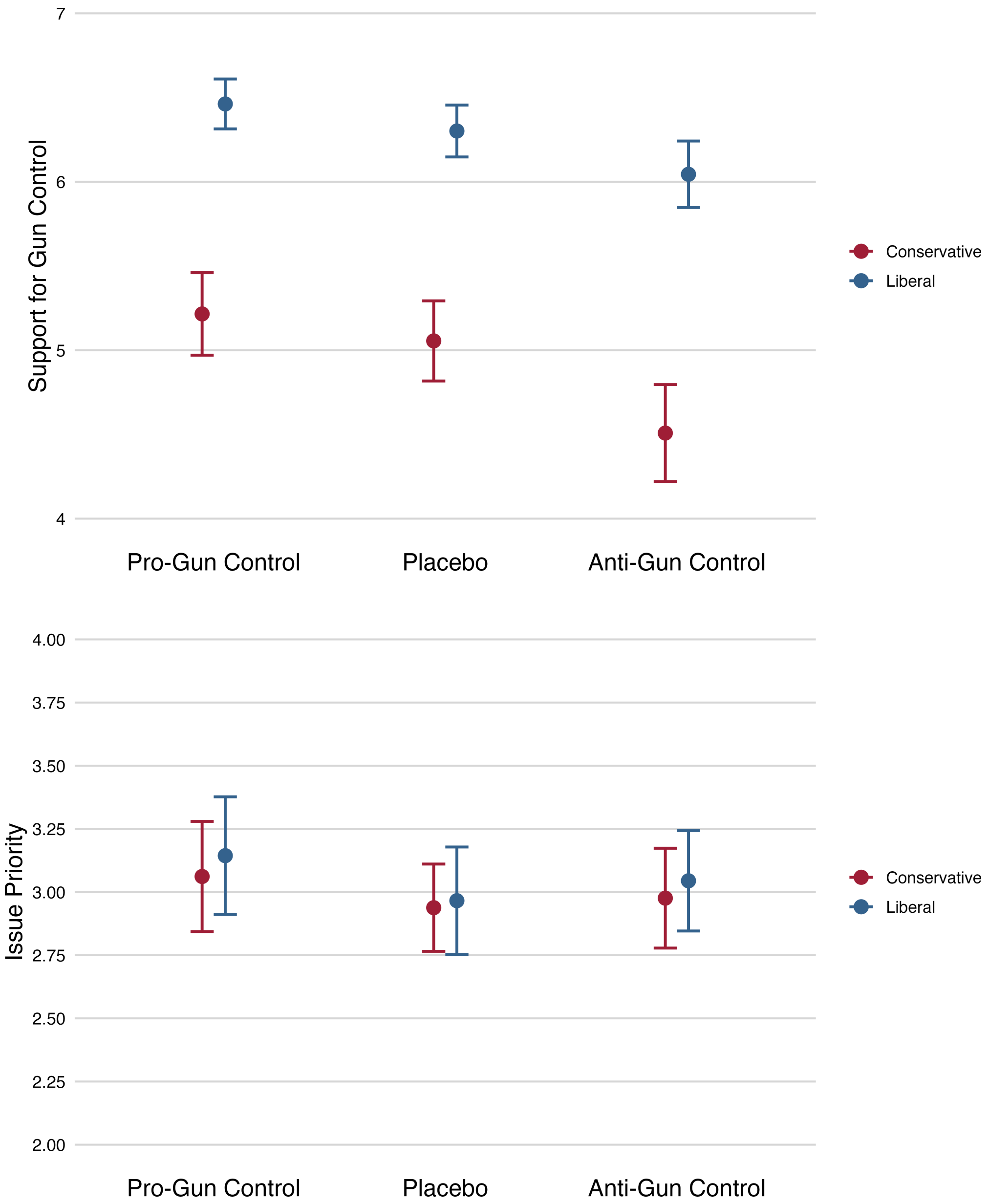
Note: This figure plots heterogeneity by a measure of self-selection into treatment, split into 2 bins (rather than 3), making this approach more comparable to PICA. Points represent means, and error bars represent 95% confidence intervals.

Figure 4: Heterogeneity by Self-Selection into Treatment (Parametric)



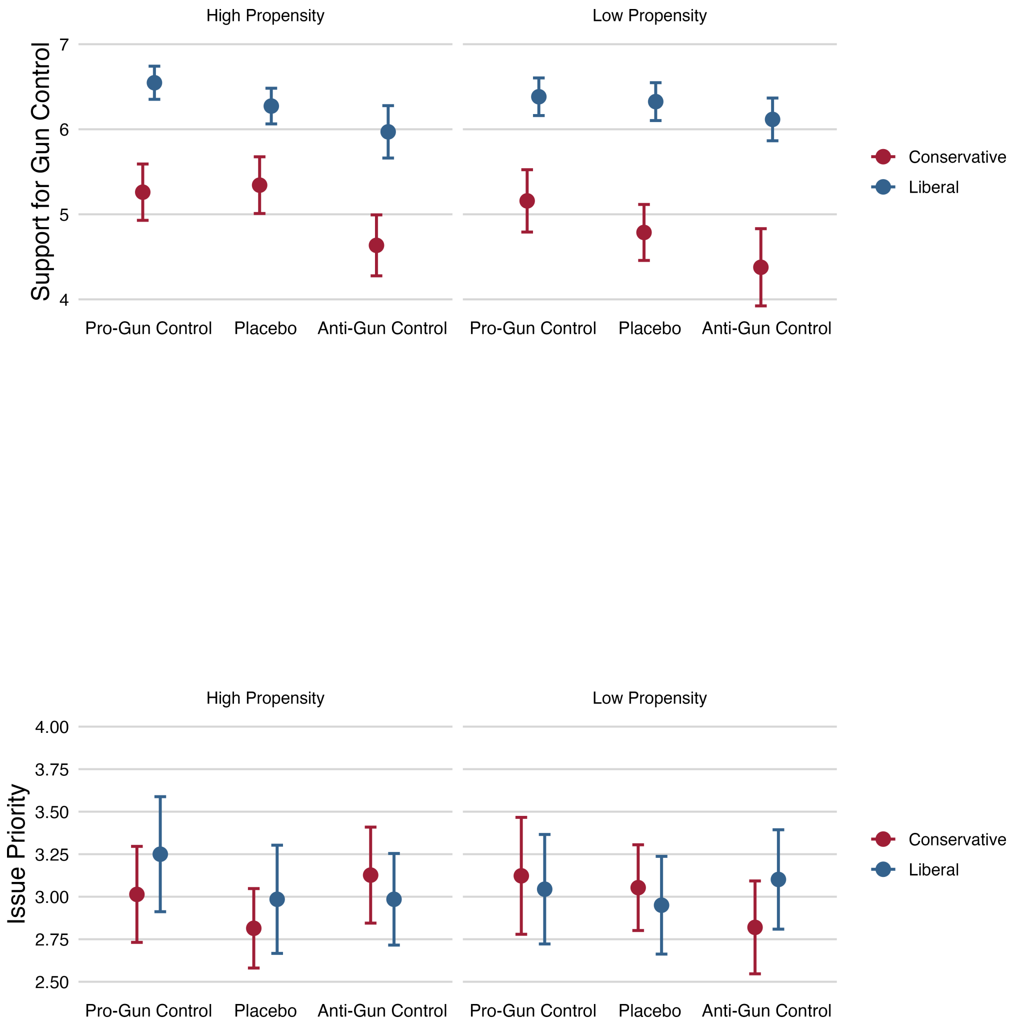
Note: This figure estimates how outcomes vary with a continuous measure of self-selection into treatment, making a second-degree polynomial functional form assumption. Points represent means, and error bars represent 95% confidence intervals.

Figure 5: Heterogeneity by Political Ideology



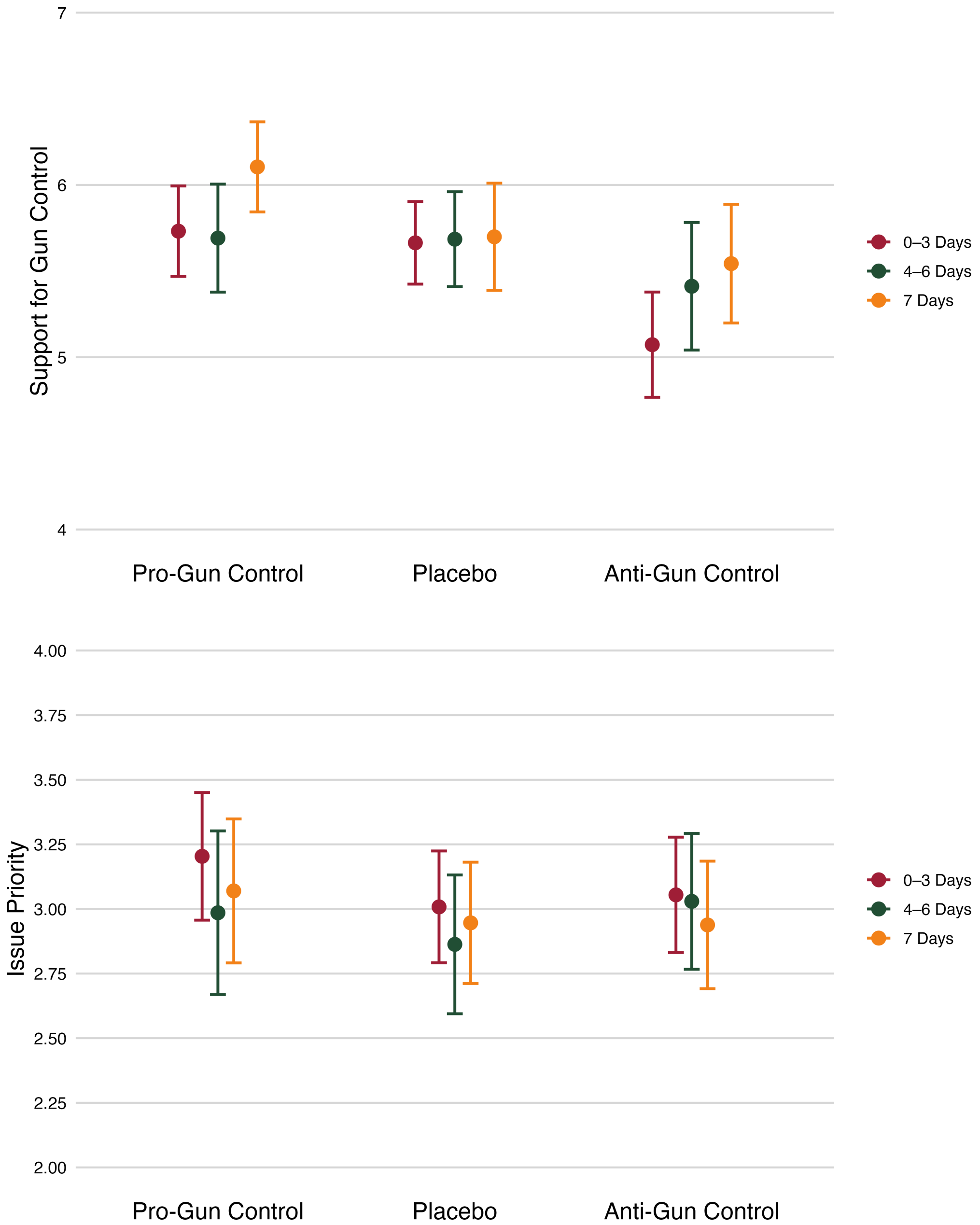
Note: This figure splits responses by treatment arm and political ideology. Points represent means and bars represent 95% confidence intervals.

Figure 6: Heterogeneity by Self-Selection and Political Ideology



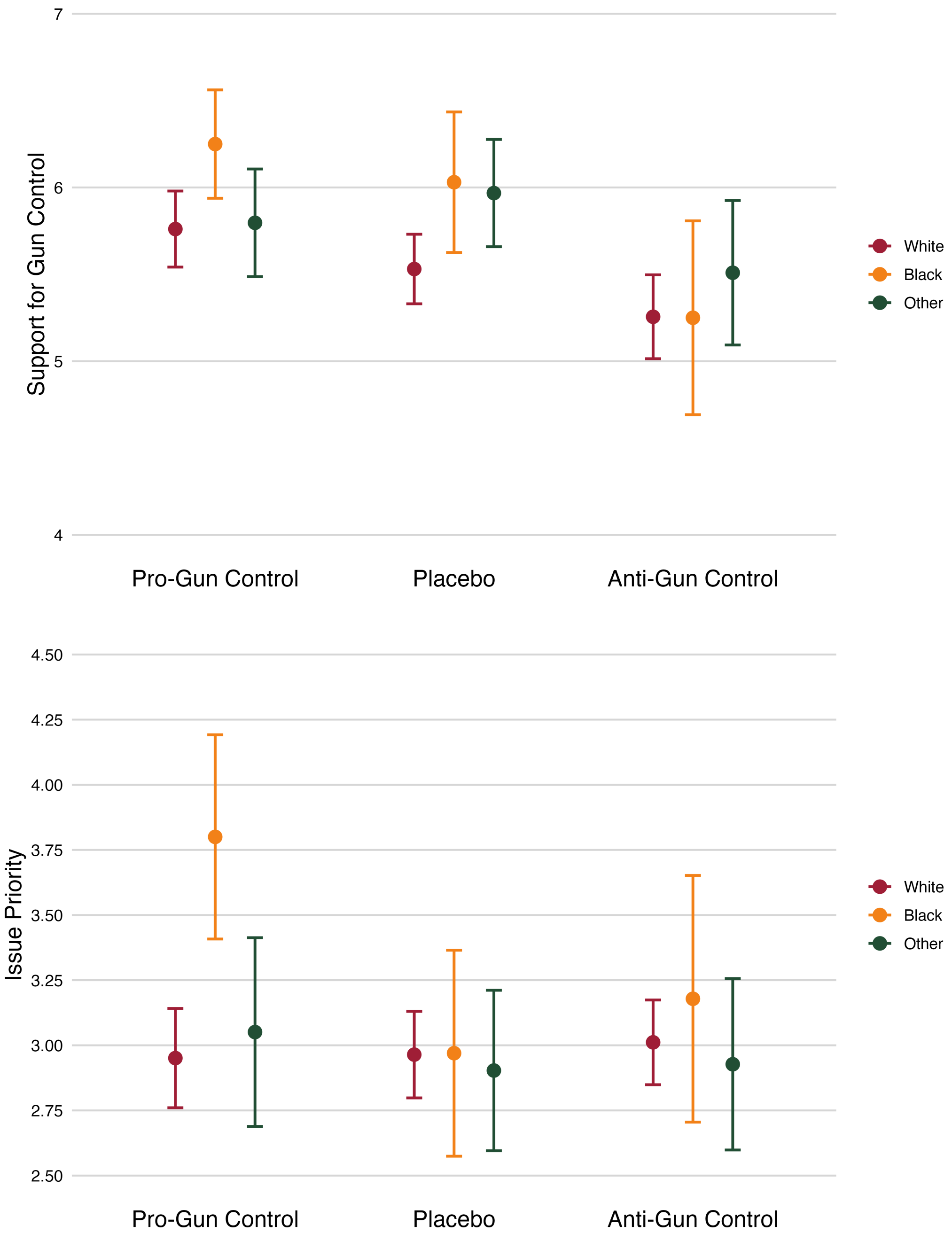
Note: This figure splits responses by treatment arm, political ideology, and propensity to self-select into treatment (in two bins). Points represent means, and bars represent 95% confidence intervals.

Figure 7: Heterogeneity by Engagement with Politics



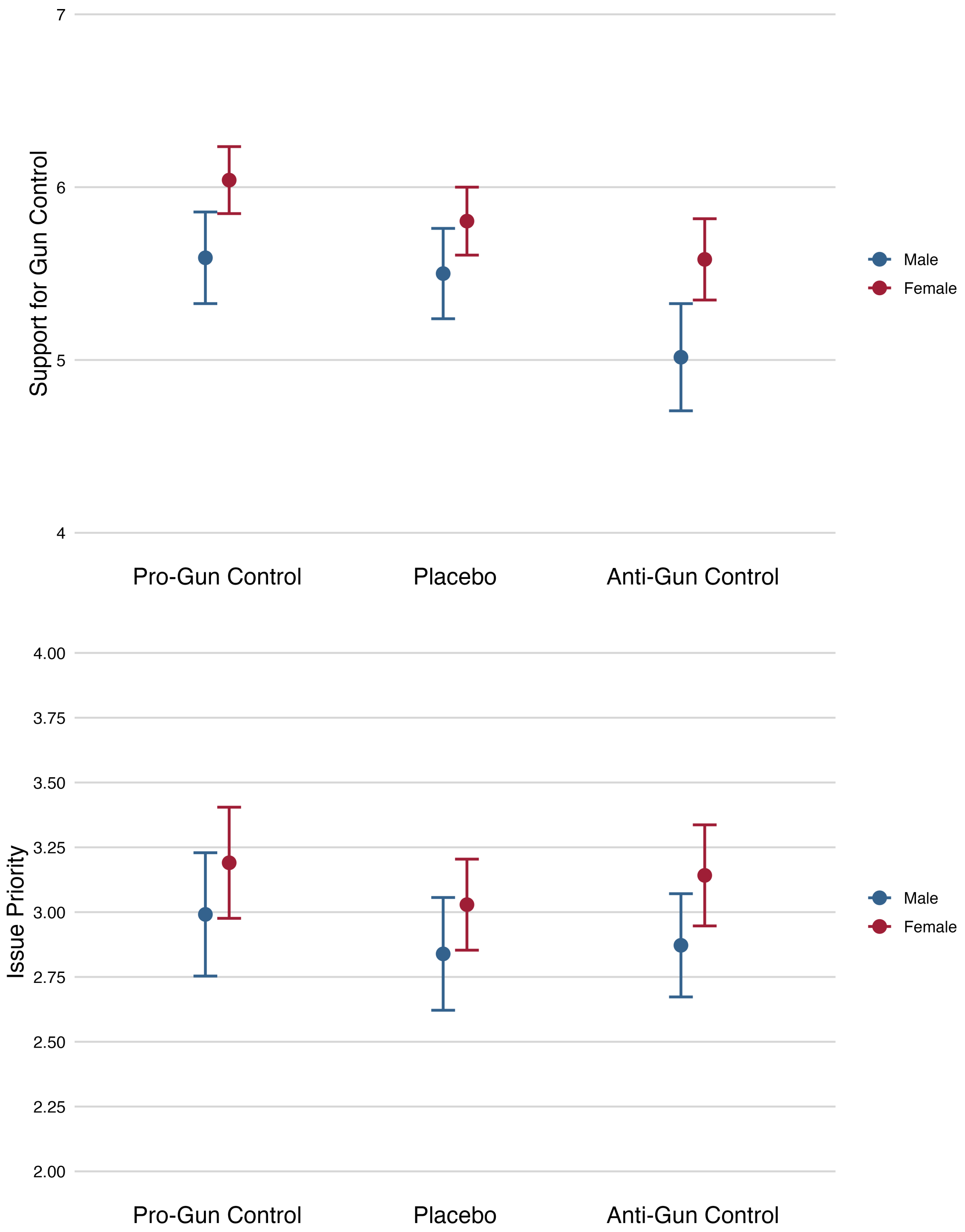
Note: This figure splits responses by treatment arm and political engagement, defined as the response to "Thinking back over the past week, how many days did you read, watch or listen to political news or opinion content?" with options (Never, One to Three Days of the Week, Four to Six Days of the Week, Every Day of the Week). Points represent means, and error bars represent 95% confidence intervals.

Figure 8: Heterogeneity by Race



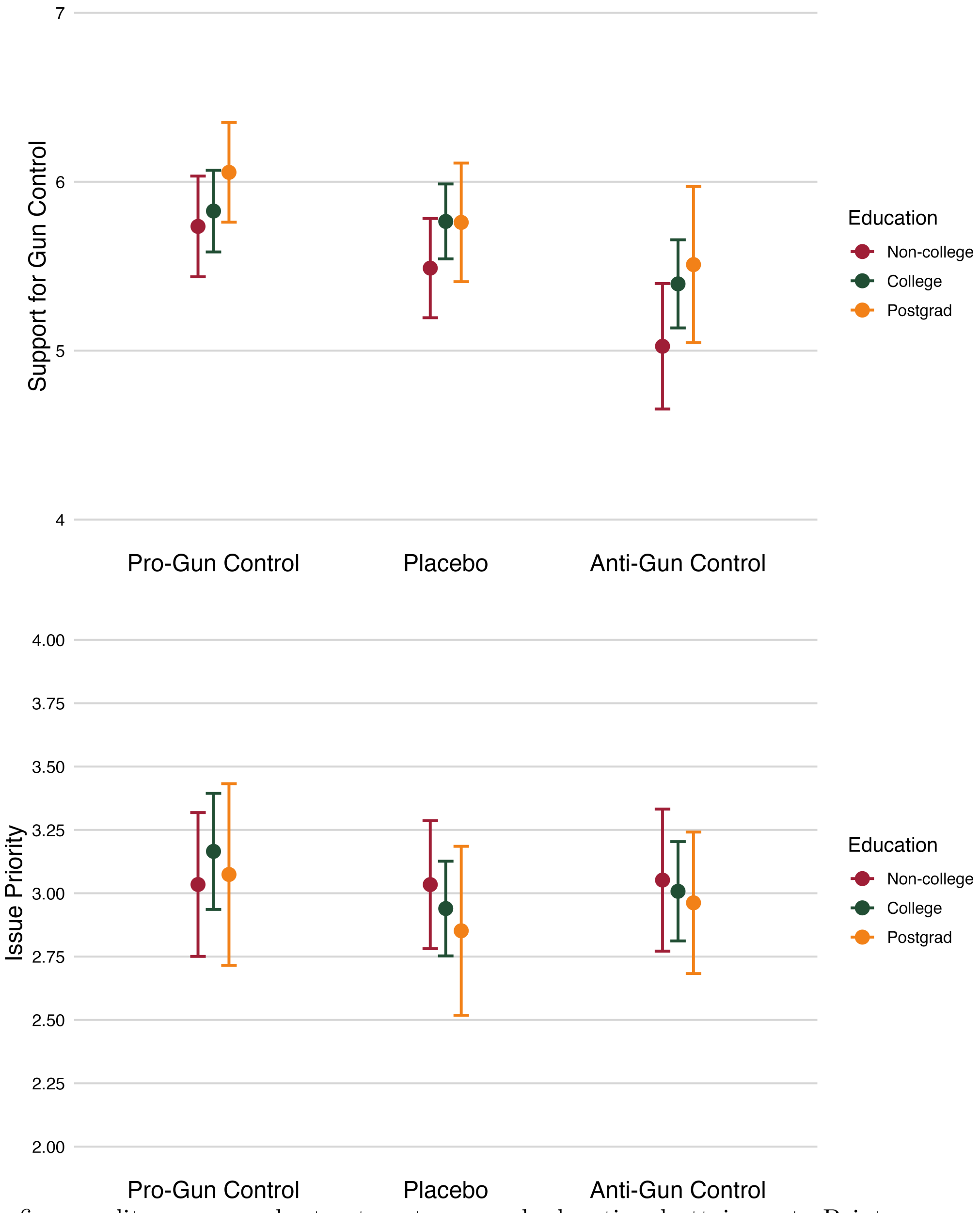
Note: This figure splits responses by treatment arm and race. Points represent means and bars represent 95% confidence intervals.

Figure 9: Heterogeneity by Gender



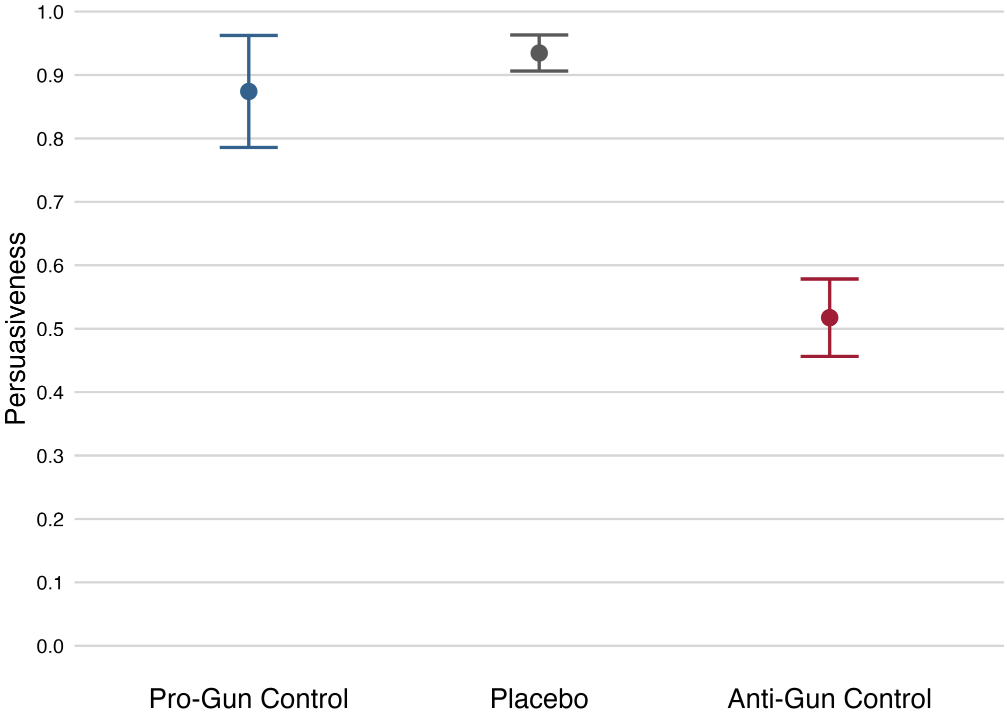
Note: This figure splits responses by treatment arm and gender. Points represent means, and error bars represent 95% confidence intervals.

Figure 10: Heterogeneity by Education



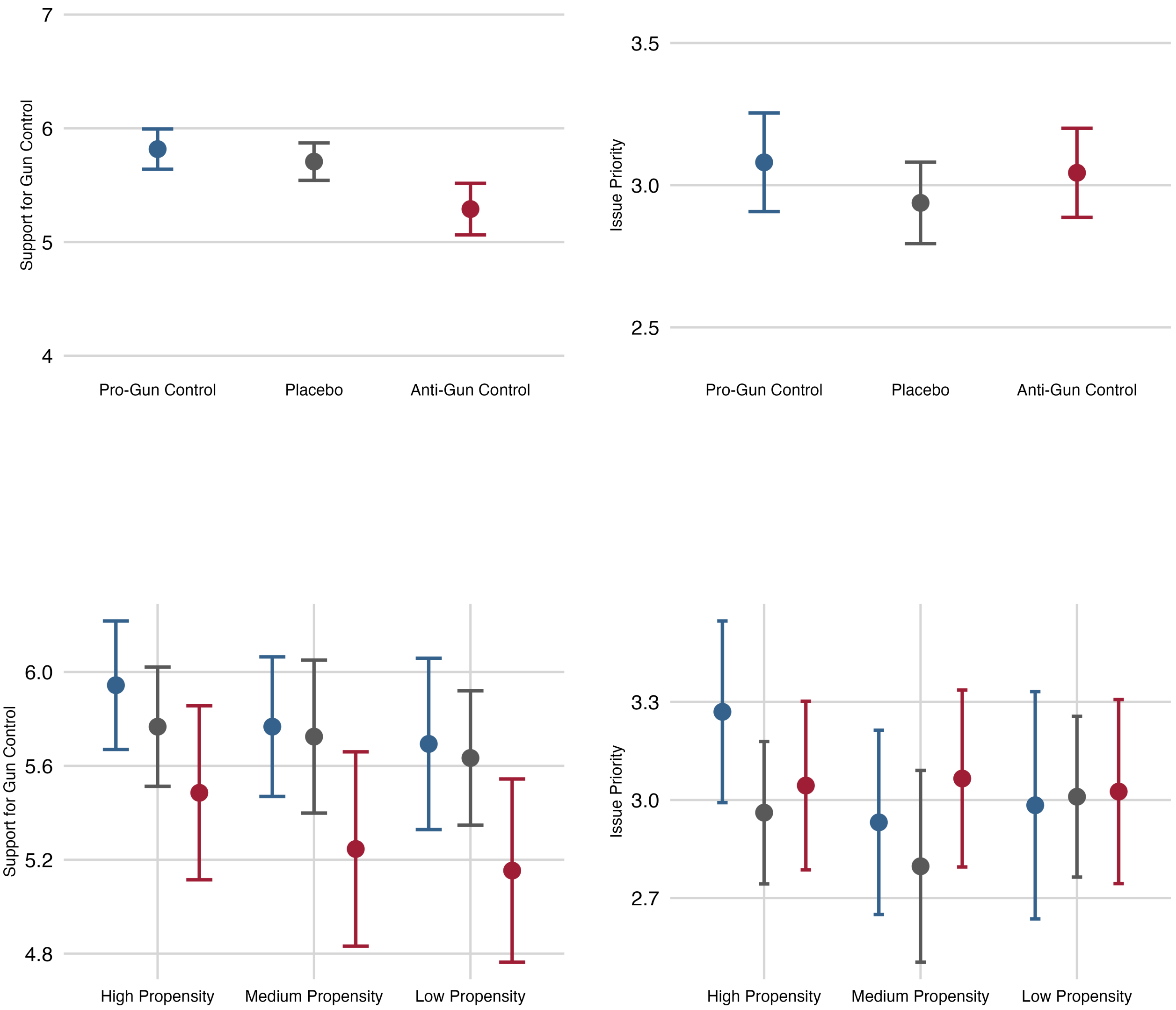
Note: This figure splits responses by treatment arm and educational attainment. Points represent means, and bars represent 95% confidence intervals.

Figure 11: Persuasiveness of Advertisements



Note: This figure graphs participant-reported persuasiveness of the ad by treatment arm. Persuasiveness is generated using gpt-4 to classify participant free responses to the question "Did you find the ad persuasive? Why or why not?", where 1 represents persuasive and 0 represents not persuasive.

Figure 12: Attention Robustness Check



Note: This figure recomputes the main figures, restricting to a sample of participants who passed two additional attention checks.

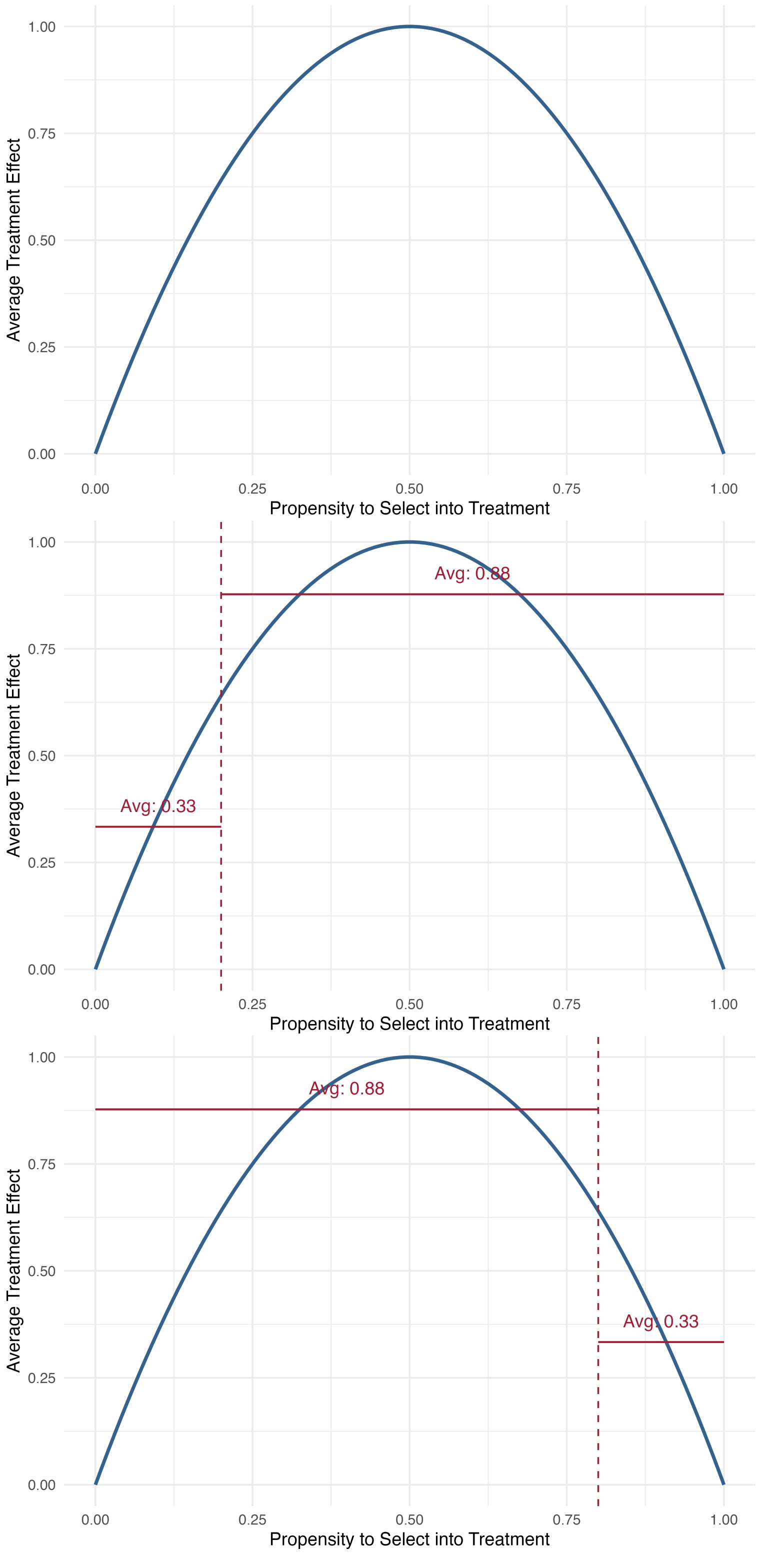


Figure 13: The first panel plots a hypothetical treatment effect curve. The second and third panels demonstrate how a PICA design can estimate different treatment effects depending on the strength of the outside option.

C Tables

Table 1: Main Regression Results

Dependent Variables:	Attitude		Issue Priority	
Model:	(1)	(2)	(3)	(4)
Variables				
Constant	5.680***	6.493***	2.952***	2.837***
	(0.0844)	(0.1716)	(0.0715)	(0.1713)
Pro-Gun Control	0.1631	0.1150	0.1512	0.1489
	(0.1227)	(0.1042)	(0.1039)	(0.1041)
Anti-Gun Control	-0.3715***	-0.3554***	0.0597	0.0790
	(0.1231)	(0.1048)	(0.1042)	(0.1046)
Controls	False	True	False	True
Fit statistics				
R^2	0.02281	0.30991	0.00263	0.02145
Observations	812	812	812	812

IID standard-errors in parentheses

*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Table 2: HTE by Self-Selection: Issue Attitude

Dependent Variable:	Attitude					
Model:	(1)	(2)	(3)	(4)	(5)	(6)
Variables						
Constant	5.748***	6.256***	5.704***	6.774***	5.596***	6.679***
	(0.1296)	(0.2910)	(0.1677)	(0.3010)	(0.1484)	(0.3053)
Anti-Gun Control	-0.3882**	-0.4216**	-0.3892*	-0.3729*	-0.3386	-0.2968*
	(0.1943)	(0.1774)	(0.2355)	(0.1980)	(0.2163)	(0.1776)
Pro-Gun Control	0.1829	0.1702	0.1253	0.1323	0.1434	-0.1174
	(0.1878)	(0.1686)	(0.2254)	(0.1876)	(0.2344)	(0.1928)
Controls	False	True	False	True	False	True
Fit statistics						
R^2	0.02786	0.25858	0.02379	0.36829	0.01607	0.37649
Observations	301	301	232	232	279	279

IID standard-errors in parentheses

*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Table 3: HTE by Self-Selection: Issue Priority

Dependent Variable:	Issue Priority					
Model:	(1)	(2)	(3)	(4)	(5)	(6)
Variables						
Constant	2.973***	2.836***	2.803***	2.701***	3.028***	3.002***
	(0.1130)	(0.2858)	(0.1437)	(0.3097)	(0.1205)	(0.3037)
Anti-Gun Control	0.0495	0.1077	0.1561	0.1754	0.0137	0.0507
	(0.1694)	(0.1743)	(0.2018)	(0.2037)	(0.1756)	(0.1767)
Pro-Gun Control	0.3340**	0.3830**	0.1290	0.1708	-0.0001	-0.0681
	(0.1637)	(0.1656)	(0.1931)	(0.1931)	(0.1902)	(0.1918)
Controls	False	True	False	True	False	True
Fit statistics						
R^2	0.01550	0.04663	0.00301	0.07021	2.75×10^{-5}	0.04812
Observations	301	301	232	232	279	279

IID standard-errors in parentheses

*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*