Online Supplementary Appendix II

Elements of External Validity:
Framework, Design, and Analysis

Table of Contents

H Statistical Details of Proposed Methodologies **1**

H.1 Contextual Exclusion Restriction 1

H.2 Proofs for Effect-Generalization 1

H.3 Proofs for Sign-Generalization 7

I Simulations **9**

I.1 Full Simulations 10

I.2 Naturalistic Simulations 15

J Literature Review of *American Political Science Review* **20**

K Numeric Results and Model Specification **26**

K.1 Results for Broockman and Kalla (2016) analysis in Figure 7 26

K.2 Results for Bisgaard (2019) analysis in Figure 8 30

K.3 Results for Broockman and Kalla (2016) analysis in Figure A1 32

K.4 Results for Bisgaard (2019) analysis in Figure A2 33

K.5 Results for Young (2019) analysis in Figure A3 35

K.6 Results for Dunning et al. (2019) analysis in Figure A4 38

K.7 Results for Dehejia, Pop-Eleches and Samii (2021) analysis in Figure A5 . . 40

K.8 Results for Bisbee et al. (2017) analysis in Figure A6 48

H Statistical Details of Proposed Methodologies

In this section, we provide proofs for all theoretical results we discussed in the paper.

H.1 Contextual Exclusion Restriction

Here, we offer a causal graphical approach to provide an alternative interpretation of contextual exclusion restriction, even though their statistical meaning is the same.

Contextual exclusion restriction can be written in a causal DAG (Figure A7). Most importantly, this causal DAG clearly shows that contextual exclusion restriction requires that variable C_i has no direct causal effect on the outcome once fixing context-moderators.

We note that in the theory of the DAG, a DAG allows for any interactions between explanatory variables to explain the outcome variable. Therefore, in the DAG (Figure A7), both C and T have a path to the outcome Y (while the effect of C is mediated by M), and thus, this mathematically means that the effect of T can be moderated by C. Therefore, the DAG shows that the causal effect will be different across contexts because C changes the causal relationship between the treatment and the outcome.

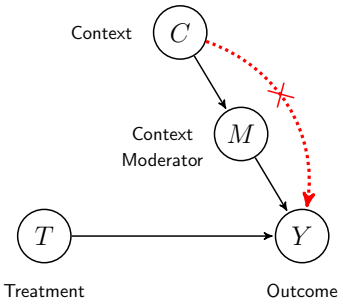

Figure A7: Causal DAG for Contextual Exclusion Restriction.

H.2 Proofs for Effect-Generalization

We examine estimation of the T-PATE when dealing with X- and C-validity together. The well-researched problem of X-validity is a special case of this setting.

H.2.1 IPW Estimator

To account for the X- and C-validity together, we need to extend conventional sampling weights that only consider the X-validity. In particular, we need two sampling weights:

$$\pi_i = \frac{1}{\Pr(S_i = 1 \mid C_i = c, \mathbf{M}_i, \mathbf{X}_i)} \qquad \text{(selection into experiments)}$$

$$\theta_i = \frac{\Pr(C_i = c^* \mid \mathbf{M}_i, \mathbf{X}_i)}{\Pr(C_i = c \mid \mathbf{M}_i, \mathbf{X}_i)} \qquad \text{(selection into contexts)}$$

Using these two sampling weights, we can show the consistency of the inverse probability weighted (IPW) estimator.

Theorem A2 (Consistency of IPW Estimator)
Consider the following IPW estimator.

$$\widehat{\tau}_{\texttt{IPW}} \equiv \frac{\sum_{i=1}^{R} \widehat{\theta}_i \widehat{\pi}_i \delta_i \mathbf{1}\{C_i = c\} S_i T_i Y_i}{\sum_{i=1}^{R} \widehat{\theta}_i \widehat{\pi}_i \delta_i \mathbf{1}\{C_i = c\} S_i T_i} - \frac{\sum_{i=1}^{R} \widehat{\theta}_i \widehat{\pi}_i (1-\delta_i) \mathbf{1}\{C_i = c\} S_i (1-T_i) Y_i}{\sum_{i=1}^{R} \widehat{\theta}_i \widehat{\pi}_i (1-\delta_i) \mathbf{1}\{C_i = c\} S_i (1-T_i)}, \tag{1}$$

where $\delta_i \equiv \Pr(T_i = 1 \mid S_i = 1, C_i = c, \mathbf{M}_i, \mathbf{X}_i)$ is the treatment assignment probability known from the experimental design. We use R to denote the sum of the sample size in the experiment (n) and in the target population data (N). Then, as $R \to \infty$,

$$\widehat{\tau}_{\texttt{IPW}} \xrightarrow{p} \mathbb{E}_{\mathcal{P}^*}[Y_i(T=1, c^*) - Y_i(T=0, c^*)], \tag{2}$$

when the sampling models are correctly specified, i.e., $\widehat{\theta}_i \xrightarrow{p} \theta_i$ and $\widehat{\pi}_i \xrightarrow{p} \pi_i$.

Proof. By the weak law of large number, $\frac{1}{R}\sum_{i=1}^{R} \widehat{\theta}_i \widehat{\pi}_i \delta_i \mathbf{1}\{C_i = c\} S_i T_i Y_i \xrightarrow{p} \mathbb{E}[\theta_i \pi_i \delta_i \mathbf{1}\{C_i = c\} S_i T_i Y_i]$ under the standard regularity conditions and the correct specification of the sampling models.

$$\begin{aligned}
& \mathbb{E}[\pi_i \theta_i \delta_i \mathbf{1}\{C_i = c\} S_i T_i Y_i] \\
= & \sum_{\mathbf{m} \in \mathcal{M}} \sum_{\mathbf{x} \in \mathcal{X}} \mathbb{E}[\pi_i \theta_i \delta_i \mathbf{1}\{C_i = c\} S_i T_i Y_i \mid \mathbf{M}_i = \mathbf{m}, \mathbf{X}_i = \mathbf{x}] \Pr(\mathbf{M}_i = \mathbf{m}, \mathbf{X}_i = \mathbf{x}) \\
= & \sum_{\mathbf{m} \in \mathcal{M}} \sum_{\mathbf{x} \in \mathcal{X}} \frac{\Pr(C_i = c^* \mid \mathbf{M}_i = \mathbf{m}, \mathbf{X}_i = \mathbf{x})}{\Pr(C_i = c \mid \mathbf{M}_i = \mathbf{m}, \mathbf{X}_i = \mathbf{x})} \\
& \qquad \times \mathbb{E}[\pi_i \delta_i \mathbf{1}\{C_i = c\} S_i T_i Y_i \mid \mathbf{M}_i = \mathbf{m}, \mathbf{X}_i = \mathbf{x}] \Pr(\mathbf{M}_i = \mathbf{m}, \mathbf{X}_i = \mathbf{x}) \\
= & \sum_{\mathbf{m} \in \mathcal{M}} \sum_{\mathbf{x} \in \mathcal{X}} \frac{\Pr(C_i = c^* \mid \mathbf{M}_i = \mathbf{m}, \mathbf{X}_i = \mathbf{x})}{\Pr(C_i = c \mid \mathbf{M}_i = \mathbf{m}, \mathbf{X}_i = \mathbf{x})} \times \Pr(C_i = c \mid \mathbf{M}_i = \mathbf{m}, \mathbf{X}_i = \mathbf{x}) \\
& \qquad \times \mathbb{E}[\pi_i \delta_i S_i T_i Y_i \mid C_i = c, \mathbf{M}_i = \mathbf{m}, \mathbf{X}_i = \mathbf{x}] \Pr(\mathbf{M}_i = \mathbf{m}, \mathbf{X}_i = \mathbf{x}) \\
= & \sum_{\mathbf{m} \in \mathcal{M}} \sum_{\mathbf{x} \in \mathcal{X}} \frac{\Pr(C_i = c^* \mid \mathbf{M}_i = \mathbf{m}, \mathbf{X}_i = \mathbf{x})}{\Pr(C_i = c \mid \mathbf{M}_i = \mathbf{m}, \mathbf{X}_i = \mathbf{x})} \times \Pr(C_i = c \mid \mathbf{M}_i = \mathbf{m}, \mathbf{X}_i = \mathbf{x}) \\
& \qquad \times \frac{1}{\Pr(S_i = 1 \mid C_i = c, \mathbf{M}_i = \mathbf{m}, \mathbf{X}_i = \mathbf{x})} \\
& \qquad \times \mathbb{E}[\delta_i S_i T_i Y_i \mid C_i = c, \mathbf{M}_i = \mathbf{m}, \mathbf{X}_i = \mathbf{x}] \Pr(\mathbf{M}_i = \mathbf{m}, \mathbf{X}_i = \mathbf{x}) \\
= & \sum_{\mathbf{m} \in \mathcal{M}} \sum_{\mathbf{x} \in \mathcal{X}} \frac{\Pr(C_i = c^* \mid \mathbf{M}_i = \mathbf{m}, \mathbf{X}_i = \mathbf{x})}{\Pr(C_i = c \mid \mathbf{M}_i = \mathbf{m}, \mathbf{X}_i = \mathbf{x})} \times \Pr(C_i = c \mid \mathbf{M}_i = \mathbf{m}, \mathbf{X}_i = \mathbf{x}) \\
& \qquad \times \frac{1}{\Pr(S_i = 1 \mid C_i = c, \mathbf{M}_i = \mathbf{m}, \mathbf{X}_i = \mathbf{x})} \times \Pr(S_i = 1 \mid C_i = c, \mathbf{M}_i = \mathbf{m}, \mathbf{X}_i = \mathbf{x}) \\
& \qquad \times \mathbb{E}[\delta_i T_i Y_i \mid S_i = 1, C_i = c, \mathbf{M}_i = \mathbf{m}, \mathbf{X}_i = \mathbf{x}] \Pr(\mathbf{M}_i = \mathbf{m}, \mathbf{X}_i = \mathbf{x}) \\
= & \sum_{\mathbf{m} \in \mathcal{M}} \sum_{\mathbf{x} \in \mathcal{X}} \frac{\Pr(C_i = c^* \mid \mathbf{M}_i = \mathbf{m}, \mathbf{X}_i = \mathbf{x})}{\Pr(C_i = c \mid \mathbf{M}_i = \mathbf{m}, \mathbf{X}_i = \mathbf{x})} \times \Pr(C_i = c \mid \mathbf{M}_i = \mathbf{m}, \mathbf{X}_i = \mathbf{x}) \\
& \qquad \times \frac{1}{\Pr(S_i = 1 \mid C_i = c, \mathbf{M}_i = \mathbf{m}, \mathbf{X}_i = \mathbf{x})} \times \Pr(S_i = 1 \mid C_i = c, \mathbf{M}_i = \mathbf{m}, \mathbf{X}_i = \mathbf{x}) \\
& \qquad \times \frac{1}{\Pr(T_i = 1 \mid S_i = 1, C_i = c, \mathbf{M}_i = \mathbf{m}, \mathbf{X}_i = \mathbf{x})} \\
& \qquad \times \mathbb{E}[T_i Y_i \mid S_i = 1, C_i = c, \mathbf{M}_i = \mathbf{m}, \mathbf{X}_i = \mathbf{x}] \Pr(\mathbf{M}_i = \mathbf{m}, \mathbf{X}_i = \mathbf{x})
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{\mathbf{m}\in\mathcal{M}}\sum_{\mathbf{x}\in\mathcal{X}} \frac{\Pr(C_i = c^* \mid \mathbf{M}_i = \mathbf{m}, \mathbf{X}_i = \mathbf{x})}{\Pr(C_i = c \mid \mathbf{M}_i = \mathbf{m}, \mathbf{X}_i = \mathbf{x})} \times \Pr(C_i = c \mid \mathbf{M}_i = \mathbf{m}, \mathbf{X}_i = \mathbf{x}) \\
&\qquad \times \frac{1}{\Pr(S_i = 1 \mid C_i = c, \mathbf{M}_i = \mathbf{m}, \mathbf{X}_i = \mathbf{x})} \times \Pr(S_i = 1 \mid C_i = c, \mathbf{M}_i = \mathbf{m}, \mathbf{X}_i = \mathbf{x}) \\
&\qquad \times \frac{1}{\Pr(T_i = 1 \mid S_i = 1, C_i = c, \mathbf{M}_i = \mathbf{m}, \mathbf{X}_i = \mathbf{x})} \times \Pr(T_i = 1 \mid S_i = 1, C_i = c, \mathbf{M}_i = \mathbf{m}, \mathbf{X}_i = \mathbf{x}) \\
&\qquad \times \mathbb{E}[Y_i(1) \mid S_i = 1, C_i = c, \mathbf{M}_i = \mathbf{m}, \mathbf{X}_i = \mathbf{x}] \Pr(\mathbf{M}_i = \mathbf{m}, \mathbf{X}_i = \mathbf{x}) \\
&= \sum_{\mathbf{m}\in\mathcal{M}}\sum_{\mathbf{x}\in\mathcal{X}} \mathbb{E}[Y_i(1) \mid S_i = 1, C_i = c, \mathbf{M}_i = \mathbf{m}, \mathbf{X}_i = \mathbf{x}] \\
&\qquad \times \Pr(C_i = c^* \mid \mathbf{M}_i = \mathbf{m}, \mathbf{X}_i = \mathbf{x}) \Pr(\mathbf{M}_i = \mathbf{m}, \mathbf{X}_i = \mathbf{x}) \\
&= \left\{ \sum_{\mathbf{m}\in\mathcal{M}}\sum_{\mathbf{x}\in\mathcal{X}} \mathbb{E}[Y_i(1) \mid \mathbf{M}_i = \mathbf{m}, \mathbf{X}_i = \mathbf{x}] \Pr(\mathbf{M}_i = \mathbf{m}, \mathbf{X}_i = \mathbf{x} \mid C_i = c^*) \right\} \Pr(C_i = c^*) \\
&= \mathbb{E}_{\mathcal{P}^*}[Y_i(T = 1, c^*)] \times \Pr(C_i = c^*).
\end{aligned}
$$

Similarly, we can show that $\frac{1}{R}\sum_{i=1}^{R} \widehat{\theta}_i\widehat{\pi}_i(1-\delta_i)\mathbf{1}\{C_i = c\}S_i(1-T_i)Y_i \xrightarrow{p} \mathbb{E}_{\mathcal{P}^*}[Y_i(T = 0, c^*)] \times \Pr(C_i = c^*)$, $\frac{1}{R}\sum_{i=1}^{R} \widehat{\theta}_i\widehat{\pi}_i\delta_i\mathbf{1}\{C_i = c\}S_iT_i \xrightarrow{p} \Pr(C_i = c^*)$, and $\frac{1}{R}\sum_{i=1}^{R} \widehat{\theta}_i\widehat{\pi}_i(1-\delta_i)\mathbf{1}\{C_i = c\}S_i(1-T_i) \xrightarrow{p} \Pr(C_i = c^*)$. This completes the proof. □

H.2.2 Weighted Least Squares

Theorem A3 (Consistency of Weighted Least Squares Estimator)
Consider the following weighted least squares estimator.

$$
(\widehat{\alpha}, \widehat{\tau}_{\texttt{wLS}}, \widehat{\gamma}) = \operatorname*{argmin}_{\alpha,\tau,\gamma} \frac{1}{n}\sum_{i=1}^{n} w_i(Y_i - \alpha - \tau T_i - \mathbf{Z}_i^\top\gamma)^2 \tag{3}
$$

where $w_i = \widehat{\theta}_i\widehat{\pi}_i\{\delta_iT_i + (1-\delta_i)(1-T_i)\}$, and $\mathbf{Z}_i$ are pre-treatment covariates measured within the experiment. Then, as $n \to \infty$,

$$
\widehat{\tau}_{\texttt{wLS}} \xrightarrow{p} \mathbb{E}_{\mathcal{P}^*}[Y_i(T = 1, c^*) - Y_i(T = 0, c^*)], \tag{4}
$$

when the sampling models are correctly specified, i.e., $\widehat{\theta}_i \xrightarrow{p} \theta_i$ and $\widehat{\pi}_i \xrightarrow{p} \pi_i$.

Proof. We rely on the proof technique by Lin (2013). Using the estimated coefficient $\widehat{\gamma}$, we can rewrite the main objective function as follows.

$$
(\widehat{\alpha}, \widehat{\tau}_{\texttt{wLS}}) = \operatorname*{argmin}_{\alpha,\tau,\gamma} \frac{1}{n}\sum_{i=1}^{n} w_i\{(Y_i - \mathbf{Z}_i^\top\widehat{\gamma}) - \alpha - \tau T_i\}^2.
$$

Therefore, using the well-known equivalence between the weighted least squares regression and the weighted difference-in-means, we can write that

$$
\widehat{\tau}_{\texttt{wLS}} = \frac{\sum_{i=1}^{n} w_i(Y_i - \mathbf{Z}_i^\top\widehat{\gamma})T_i}{\sum_{i=1}^{n} w_iT_i} - \frac{\sum_{i=1}^{n} w_i(Y_i - \mathbf{Z}_i^\top\widehat{\gamma})(1-T_i)}{\sum_{i=1}^{n} w_i(1-T_i)}.
$$

We now further examine this quantity.

$$
\widehat{\tau}_{\texttt{wLS}} = \frac{\sum_{i=1}^{n} w_i(Y_i - \mathbf{Z}_i^\top\widehat{\gamma})T_i}{\sum_{i=1}^{n} w_iT_i} - \frac{\sum_{i=1}^{n} w_i(Y_i - \mathbf{Z}_i^\top\widehat{\gamma})(1-T_i)}{\sum_{i=1}^{n} w_i(1-T_i)}
$$

$$= \frac{\sum_{i=1}^n w_i Y_i T_i}{\sum_{i=1}^n w_i T_i} - \frac{\sum_{i=1}^n w_i Y_i (1-T_i)}{\sum_{i=1}^n w_i (1-T_i)} + \left\{ \frac{\sum_{i=1}^n w_i \mathbf{Z}_i^\top T_i}{\sum_{i=1}^n w_i T_i} - \frac{\sum_{i=1}^n w_i \mathbf{Z}_i^\top (1-T_i)}{\sum_{i=1}^n w_i (1-T_i)} \right\} \widehat{\gamma}.$$

Using the weak law of large number, $\frac{1}{n}\sum_{i=1}^n w_i \mathbf{Z}_i^\top T_i \xrightarrow{p} \mathbb{E}[w_i \mathbf{Z}_i^\top T_i]$, $\frac{1}{n}\sum_{i=1}^n w_i \mathbf{Z}_i^\top (1-T_i) \xrightarrow{p} \mathbb{E}[w_i \mathbf{Z}_i^\top (1-T_i)]$, $\frac{1}{n}\sum_{i=1}^n w_i T_i \xrightarrow{p} \mathbb{E}[w_i T_i]$, and $\frac{1}{n}\sum_{i=1}^n w_i (1-T_i) \xrightarrow{p} \mathbb{E}[w_i (1-T_i)]$.

We can also show that

$$\begin{aligned}
\mathbb{E}[w_i \mathbf{Z}_i^\top T_i] &= \sum_{\mathbf{m}\in\mathcal{M}} \sum_{\mathbf{x}\in\mathcal{X}} \theta_i \pi_i \mathbb{E}[Z_i \mid S_i = 1, C_i = c, \mathbf{M}_i = \mathbf{m}, \mathbf{X}_i = \mathbf{x}]^\top \Pr(\mathbf{M}_i = \mathbf{m}, \mathbf{X}_i = \mathbf{x} \mid S_i = 1, C_i = c) \\
\mathbb{E}[w_i \mathbf{Z}_i^\top (1-T_i)] &= \sum_{\mathbf{m}\in\mathcal{M}} \sum_{\mathbf{x}\in\mathcal{X}} \theta_i \pi_i \mathbb{E}[Z_i \mid S_i = 1, C_i = c, \mathbf{M}_i = \mathbf{m}, \mathbf{X}_i = \mathbf{x}]^\top \Pr(\mathbf{M}_i = \mathbf{m}, \mathbf{X}_i = \mathbf{x} \mid S_i = 1, C_i = c) \\
\mathbb{E}[w_i T_i] &= \sum_{\mathbf{m}\in\mathcal{M}} \sum_{\mathbf{x}\in\mathcal{X}} \theta_i \pi_i \Pr(\mathbf{M}_i = \mathbf{m}, \mathbf{X}_i = \mathbf{x} \mid S_i = 1, C_i = c) \\
\mathbb{E}[w_i (1-T_i)] &= \sum_{\mathbf{m}\in\mathcal{M}} \sum_{\mathbf{x}\in\mathcal{X}} \theta_i \pi_i \Pr(\mathbf{M}_i = \mathbf{m}, \mathbf{X}_i = \mathbf{x} \mid S_i = 1, C_i = c).
\end{aligned}$$

Combined together,

$$\left\{ \frac{\sum_{i=1}^n w_i \mathbf{Z}_i^\top T_i}{\sum_{i=1}^n w_i T_i} - \frac{\sum_{i=1}^n w_i \mathbf{Z}_i^\top (1-T_i)}{\sum_{i=1}^n w_i (1-T_i)} \right\} \widehat{\gamma} \xrightarrow{p} 0,$$

given that $\widehat{\gamma}$ converges to some constant γ^* under the standard regularity conditions.

Finally, we note that

$$\begin{aligned}
& \frac{\sum_{i=1}^n w_i Y_i T_i}{\sum_{i=1}^n w_i T_i} - \frac{\sum_{i=1}^n w_i Y_i (1-T_i)}{\sum_{i=1}^n w_i (1-T_i)} \\
= \; & \frac{\sum_{i=1}^R \widehat{\theta}_i \widehat{\pi}_i \delta_i \mathbf{1}\{C_i = c\} S_i T_i Y_i}{\sum_{i=1}^R \widehat{\theta}_i \widehat{\pi}_i \delta_i \mathbf{1}\{C_i = c\} S_i T_i} - \frac{\sum_{i=1}^R \widehat{\theta}_i \widehat{\pi}_i (1-\delta_i) \mathbf{1}\{C_i = c\} S_i (1-T_i) Y_i}{\sum_{i=1}^R \widehat{\theta}_i \widehat{\pi}_i (1-\delta_i) \mathbf{1}\{C_i = c\} S_i (1-T_i)} \\
= \; & \widehat{\tau}_{\texttt{IPW}} \\
\xrightarrow{p} \; & \mathbb{E}_{\mathcal{P}^*}[Y_i(T=1, c^*) - Y_i(T=0, c^*)].
\end{aligned}$$

where we use Theorem A2. This completes the proof. □

H.2.3 Outcome-Based Estimator

Theorem A4 (Consistency of Outcome-based Estimator)
Consider the following weighted least squares estimator.

$$\widehat{\tau}_{\texttt{out}} = \frac{1}{N} \sum_{j \in \mathcal{P}^*} \{\widehat{g}_1(\mathbf{X}_j, \mathbf{M}_j) - \widehat{g}_1(\mathbf{X}_j, \mathbf{M}_j)\}$$

where

$$\begin{aligned}
\widehat{g}_1(\mathbf{X}_j, \mathbf{M}_j) &\equiv \widehat{\mathbb{E}}(Y_i \mid T_i = 1, \mathbf{M}_j, \mathbf{X}_j, S_i = 1, C_i = c), \\
\widehat{g}_0(\mathbf{X}_j, \mathbf{M}_j) &\equiv \widehat{\mathbb{E}}(Y_i \mid T_i = 0, \mathbf{M}_j, \mathbf{X}_j, S_i = 1, C_i = c).
\end{aligned}$$

Then, as $N \to \infty$,

$$\widehat{\tau}_{\texttt{out}} \xrightarrow{p} \mathbb{E}_{\mathcal{P}^*}[Y_i(T=1, c^*) - Y_i(T=0, c^*)], \tag{5}$$

when the outcome models are correctly specified, i.e., $\widehat{g}_1(\mathbf{x}, \mathbf{m}) \xrightarrow{p} \mathbb{E}(Y_i \mid T_i = 1, \mathbf{m}, \mathbf{x}, S_i = 1, C_i = c)$, and $\widehat{g}_0(\mathbf{x}, \mathbf{m}) \xrightarrow{p} \mathbb{E}(Y_i \mid T_i = 0, \mathbf{m}, \mathbf{x}, S_i = 1, C_i = c)$.

Proof. Due to the weak law of large number,

$$\begin{aligned}
&\frac{1}{N}\sum_{j\in\mathcal{P}^*}\{\widehat{g}_1(\mathbf{X}_j,\mathbf{M}_j)-\widehat{g}_1(\mathbf{X}_j,\mathbf{M}_j)\}\\
\xrightarrow{p}\;&\sum_{\mathbf{m}\in\mathcal{M}}\sum_{\mathbf{x}\in\mathcal{X}}\{\mathbb{E}(Y_i\mid T_i=1,\mathbf{M}_i=\mathbf{m},\mathbf{X}_i=\mathbf{x},S_i=1,C_i=c)-\mathbb{E}(Y_i\mid T_i=0,\mathbf{M}_i=\mathbf{m},\mathbf{X}_i=\mathbf{x},S_i=1,C_i=c)\}\\
&\qquad\times\Pr(\mathbf{M}_i=\mathbf{m},\mathbf{X}_i=\mathbf{x}\mid C_i=c^*)\\
=\;&\mathbb{E}_{\mathcal{P}^*}[Y_i(T=1,c^*)-Y_i(T=0,c^*)],
\end{aligned}$$

where we used Theorem A1 in the final equality. This completes the proof. □

H.2.4 Doubly Robust Estimator

To account for potential model misspecification, we explicitly parameterize the outcome models and sampling weights. First, we define outcome models with a finite dimensional parameter ξ; $g_1(\mathbf{M}_i,\mathbf{X}_i;\xi_1)$ and $g_1(\mathbf{M}_i,\mathbf{X}_i;\xi_0)$. We use ξ^* to denote correctly specified outcome models, that is,

$$g_1(\mathbf{M}_i,\mathbf{X}_i;\xi_1^*)=\mathbb{E}(Y_i\mid T_i=1,\mathbf{M}_i,\mathbf{X}_i,S_i=1,C_i=c), \tag{6}$$

$$g_0(\mathbf{M}_i,\mathbf{X}_i;\xi_0^*)=\mathbb{E}(Y_i\mid T_i=0,\mathbf{M}_i,\mathbf{X}_i,S_i=1,C_i=c). \tag{7}$$

Similarly, we define sampling weights as a finite dimensional parameter ψ; $\theta(\mathbf{M}_i,\mathbf{X}_i;\psi_C)$ and $\pi(\mathbf{M}_i,\mathbf{X}_i;\psi_S)$. We use ψ^* to denote correctly specified sampling weights, that is,

$$\pi(\mathbf{M}_i,\mathbf{X}_i;\psi_S^*)=\frac{1}{\Pr(S_i=1\mid C_i=c,\mathbf{M}_i,\mathbf{X}_i)}$$

$$\theta(\mathbf{M}_i,\mathbf{X}_i;\psi_C^*)=\frac{\Pr(C_i=c^*\mid\mathbf{M}_i,\mathbf{X}_i)}{\Pr(C_i=c\mid\mathbf{M}_i,\mathbf{X}_i)}$$

Then, using these extended sampling weights and outcome models, we propose the augmented IPW (AIPW) estimator as follows.

Theorem A5 (Double Robustness of AIPW Estimator)
Consider the following AIPW estimator.

$$\begin{aligned}
\widehat{\tau}_{\texttt{AIPW}}\;\equiv\;&\frac{\sum_{i=1}^R\theta(\mathbf{M}_i,\mathbf{X}_i;\widehat{\psi}_C)\pi(\mathbf{M}_i,\mathbf{X}_i;\widehat{\psi}_S)\delta_i\mathbf{1}\{C_i=c\}S_iT_i\{Y_i-g_1(\mathbf{M}_i,\mathbf{X}_i;\widehat{\xi}_1)\}}{\sum_{i=1}^R\theta(\mathbf{M}_i,\mathbf{X}_i;\widehat{\psi}_C)\pi(\mathbf{M}_i,\mathbf{X}_i;\widehat{\psi}_S)\delta_i\mathbf{1}\{C_i=c\}S_iT_i}\\
&-\frac{\sum_{i=1}^R\theta(\mathbf{M}_i,\mathbf{X}_i;\widehat{\psi}_C)\pi(\mathbf{M}_i,\mathbf{X}_i;\widehat{\psi}_S)(1-\delta_i)\mathbf{1}\{C_i=c\}S_i(1-T_i)\{Y_i-g_0(\mathbf{M}_i,\mathbf{X}_i;\widehat{\xi}_0)\}}{\sum_{i=1}^R\theta(\mathbf{M}_i,\mathbf{X}_i;\widehat{\psi}_C)\pi(\mathbf{M}_i,\mathbf{X}_i;\widehat{\psi}_S)(1-\delta_i)\mathbf{1}\{C_i=c\}S_i(1-T_i)}\\
&+\frac{\sum_{i=1}^R\mathbf{1}\{C_i=c^*\}\{g_1(\mathbf{M}_i,\mathbf{X}_i;\widehat{\xi}_1)-g_0(\mathbf{M}_i,\mathbf{X}_i;\widehat{\xi}_0)\}}{\sum_{i=1}^R\mathbf{1}\{C_i=c^*\}},
\end{aligned}$$

where we use R to denote the sum of the sample size in the experiment (n) and in the target population data (N). Then, if the outcome models or sampling weights are correctly specified, the AIPW estimator is consistent. Formally,

$$\begin{aligned}
\text{if}\quad&\{\widehat{\xi}_1\xrightarrow{p}\xi_1^*,\;\text{ and }\;\widehat{\xi}_0\xrightarrow{p}\xi_0^*\}\;\text{ or }\;\{\widehat{\psi}_C\xrightarrow{p}\psi_C^*,\;\text{ and }\;\widehat{\psi}_S\xrightarrow{p}\psi_S^*\},\\
&\widehat{\tau}_{\texttt{AIPW}}\xrightarrow{p}\mathbb{E}_{\mathcal{P}^*}[Y_i(T=1,c^*)-Y_i(T=0,c^*)],\qquad\text{as}\quad R\to\infty.
\end{aligned}$$

Proof. Following the standard convention (Tsiatis, 2006), we assume that $\widehat{\xi}_1$ and $\widehat{\xi}_0$ converge in probability to some value $\tilde{\xi}_1$ and $\tilde{\xi}_0$ as R goes to infinity. When $\tilde{\xi}_1 = \xi_1^*$ and $\tilde{\xi}_0 = \xi_0^*$, we will say the outcome models are correctly specified. Similarly, $\widehat{\psi}_C$ and $\widehat{\psi}_S$ converge in probability to some value $\tilde{\psi}_C$ and $\tilde{\psi}_S$ as R goes to infinity. When $\tilde{\psi}_C = \psi_C^*$ and $\tilde{\psi}_S = \psi_S^*$, we will say the sampling weights are correctly specified.

First, we consider cases under which sampling weights are correctly specified. Then, based on Theorem A2, we know that

$$\frac{\sum_{i=1}^R \theta(\mathbf{M}_i, \mathbf{X}_i; \widehat{\psi}_C)\pi(\mathbf{M}_i, \mathbf{X}_i; \widehat{\psi}_S)\delta_i \mathbf{1}\{C_i = c\} S_i T_i Y_i}{\sum_{i=1}^R \theta(\mathbf{M}_i, \mathbf{X}_i; \widehat{\psi}_C)\pi(\mathbf{M}_i, \mathbf{X}_i; \widehat{\psi}_S)\delta_i \mathbf{1}\{C_i = c\} S_i T_i} \tag{8}$$

$$- \frac{\sum_{i=1}^R \theta(\mathbf{M}_i, \mathbf{X}_i; \widehat{\psi}_C)\pi(\mathbf{M}_i, \mathbf{X}_i; \widehat{\psi}_S)(1-\delta_i) \mathbf{1}\{C_i = c\} S_i (1-T_i) Y_i}{\sum_{i=1}^R \theta(\mathbf{M}_i, \mathbf{X}_i; \widehat{\psi}_C)\pi(\mathbf{M}_i, \mathbf{X}_i; \widehat{\psi}_S)\delta_i \mathbf{1}\{C_i = c\} S_i (1-T_i)} \tag{9}$$

$$\xrightarrow{p} \quad \mathbb{E}_{\mathcal{P}^*}[Y_i(T=1, c^*) - Y_i(T=0, c^*)]. \tag{10}$$

Therefore, we need to verify that

$$\frac{\sum_{i=1}^R \theta(\mathbf{M}_i, \mathbf{X}_i; \widehat{\psi}_C)\pi(\mathbf{M}_i, \mathbf{X}_i; \widehat{\psi}_S)\delta_i \mathbf{1}\{C_i = c\} S_i T_i g_1(\mathbf{M}_i, \mathbf{X}_i; \widehat{\xi}_1)}{\sum_{i=1}^R \theta(\mathbf{M}_i, \mathbf{X}_i; \widehat{\psi}_C)\pi(\mathbf{M}_i, \mathbf{X}_i; \widehat{\psi}_S)\delta_i \mathbf{1}\{C_i = c\} S_i T_i}$$

$$- \frac{\sum_{i=1}^R \mathbf{1}\{C_i = c^*\} g_1(\mathbf{M}_i, \mathbf{X}_i; \widehat{\xi}_1)}{\sum_{i=1}^R \mathbf{1}\{C_i = c^*\}} \xrightarrow{p} 0$$

$$\frac{\sum_{i=1}^R \theta(\mathbf{M}_i, \mathbf{X}_i; \widehat{\psi}_C)\pi(\mathbf{M}_i, \mathbf{X}_i; \widehat{\psi}_S)(1-\delta_i)\mathbf{1}\{C_i = c\} S_i (1-T_i) g_0(\mathbf{M}_i, \mathbf{X}_i; \widehat{\xi}_0)}{\sum_{i=1}^R \theta(\mathbf{M}_i, \mathbf{X}_i; \widehat{\psi}_C)\pi(\mathbf{M}_i, \mathbf{X}_i; \widehat{\psi}_S)(1-\delta_i)\mathbf{1}\{C_i = c\} S_i (1-T_i)}$$

$$- \frac{\sum_{i=1}^R \mathbf{1}\{C_i = c^*\} g_0(\mathbf{M}_i, \mathbf{X}_i; \widehat{\xi}_0)}{\sum_{i=1}^R \mathbf{1}\{C_i = c^*\}} \xrightarrow{p} 0.$$

Using the weak law of large number, we obtain

$$\frac{1}{R}\sum_{i=1}^R \theta(\mathbf{M}_i, \mathbf{X}_i; \widehat{\psi}_C)\pi(\mathbf{M}_i, \mathbf{X}_i; \widehat{\psi}_S)\delta_i \mathbf{1}\{C_i = c\} S_i T_i g_1(\mathbf{M}_i, \mathbf{X}_i; \widehat{\xi}_1) \xrightarrow{p} \mathbb{E}[\theta_i \pi_i \delta_i \mathbf{1}\{C_i = c\} S_i T_i g_1(\mathbf{M}_i, \mathbf{X}_i; \tilde{\xi}_1)], \tag{11}$$

where we use θ_i and π_i to denote the correctly specified weights. Using the same proof strategy as Theorem A2,

$$\mathbb{E}[\theta_i \pi_i \delta_i \mathbf{1}\{C_i = c\} S_i T_i g_1(\mathbf{M}_i, \mathbf{X}_i; \tilde{\xi}_1)] \tag{12}$$

$$= \sum_{\mathbf{m}\in\mathcal{M}} \sum_{\mathbf{x}\in\mathcal{X}} \Big\{ g_1(\mathbf{M}_i, \mathbf{X}_i; \tilde{\xi}_1) \Pr(\mathbf{M}_i = \mathbf{m}, \mathbf{X}_i = \mathbf{x} \mid C_i = c^*) \Big\} \Pr(C_i = c^*). \tag{13}$$

Therefore, we get

$$\frac{\sum_{i=1}^R \theta(\mathbf{M}_i, \mathbf{X}_i; \widehat{\psi}_C)\pi(\mathbf{M}_i, \mathbf{X}_i; \widehat{\psi}_S)\delta_i \mathbf{1}\{C_i = c\} S_i T_i g_1(\mathbf{M}_i, \mathbf{X}_i; \widehat{\xi}_1)}{\sum_{i=1}^R \theta(\mathbf{M}_i, \mathbf{X}_i; \widehat{\psi}_C)\pi(\mathbf{M}_i, \mathbf{X}_i; \widehat{\psi}_S)\delta_i \mathbf{1}\{C_i = c\} S_i T_i} \xrightarrow{p} \mathbb{E}\Big\{ g_1(\mathbf{M}_i, \mathbf{X}_i; \tilde{\xi}_1) \mid C_i = c^* \Big\}$$

It is easy to see that $\dfrac{\sum_{i=1}^R \mathbf{1}\{C_i = c^*\} g_1(\mathbf{M}_i, \mathbf{X}_i; \widehat{\xi}_1)}{\sum_{i=1}^R \mathbf{1}\{C_i = c^*\}} \xrightarrow{p} \mathbb{E}\Big\{ g_1(\mathbf{M}_i, \mathbf{X}_i; \tilde{\xi}_1) \mid C_i = c^* \Big\}$. We can use the similar proof for the expression for the control group. Thus, we obtain the desired result for cases when sampling weights are correctly specified.

Next, we consider cases under which the outcome models are correctly specified. In this case, we have

$$\frac{\sum_{i=1}^{R} \mathbf{1}\{C_i = c^*\}\{\widehat{g}_1(\mathbf{M}_i, \mathbf{X}_i) - \widehat{g}_0(\mathbf{M}_i, \mathbf{X}_i)\}}{\sum_{i=1}^{R} \mathbf{1}\{C_i = c^*\}} \tag{14}$$

$$\xrightarrow{p} \quad \mathbb{E}_{\mathcal{P}^*}[Y_i(T=1, c^*) - Y_i(T=0, c^*)]. \tag{15}$$

Therefore, we need to verify that

$$\frac{\sum_{i=1}^{R} \theta(\mathbf{M}_i, \mathbf{X}_i; \widehat{\psi}_C)\pi(\mathbf{M}_i, \mathbf{X}_i; \widehat{\psi}_S)\delta_i \mathbf{1}\{C_i = c\} S_i T_i \{Y_i - g_1(\mathbf{M}_i, \mathbf{X}_i; \widehat{\xi}_1)\}}{\sum_{i=1}^{R} \theta(\mathbf{M}_i, \mathbf{X}_i; \widehat{\psi}_C)\pi(\mathbf{M}_i, \mathbf{X}_i; \widehat{\psi}_S)\delta_i \mathbf{1}\{C_i = c\} S_i T_i} \xrightarrow{p} 0$$

$$\frac{\sum_{i=1}^{R} \theta(\mathbf{M}_i, \mathbf{X}_i; \widehat{\psi}_C)\pi(\mathbf{M}_i, \mathbf{X}_i; \widehat{\psi}_S)(1-\delta_i) \mathbf{1}\{C_i = c\} S_i (1-T_i) \{Y_i - g_0(\mathbf{M}_i, \mathbf{X}_i; \widehat{\xi}_0)\}}{\sum_{i=1}^{R} \theta(\mathbf{M}_i, \mathbf{X}_i; \widehat{\psi}_C)\pi(\mathbf{M}_i, \mathbf{X}_i; \widehat{\psi}_S)(1-\delta_i) \mathbf{1}\{C_i = c\} S_i (1-T_i)} \xrightarrow{p} 0$$

Now, using the weak law of large number, we obtain

$$\frac{1}{R} \sum_{i=1}^{R} \sum_{i=1}^{R} \theta(\mathbf{M}_i, \mathbf{X}_i; \widehat{\psi}_C)\pi(\mathbf{M}_i, \mathbf{X}_i; \widehat{\psi}_S)\delta_i \mathbf{1}\{C_i = c\} S_i T_i \{Y_i - g_1(\mathbf{M}_i, \mathbf{X}_i; \widehat{\xi}_1)\}$$

$$\xrightarrow{p} \quad \mathbb{E}[\theta(\mathbf{M}_i, \mathbf{X}_i; \tilde{\psi}_C)\pi(\mathbf{M}_i, \mathbf{X}_i; \tilde{\psi}_S)\delta_i \mathbf{1}\{C_i = c\} S_i T_i \{Y_i - g_1(\mathbf{M}_i, \mathbf{X}_i; \psi_1^*)\}].$$

Now, we can rewrite the expression as follows.

$$\begin{aligned}
& \mathbb{E}[\theta(\mathbf{M}_i, \mathbf{X}_i; \tilde{\psi}_C)\pi(\mathbf{M}_i, \mathbf{X}_i; \tilde{\psi}_S)\delta_i \mathbf{1}\{C_i = c\} S_i T_i \{Y_i - g_1(\mathbf{M}_i, \mathbf{X}_i; \psi_1^*)\}] \\
= & \sum_{\mathbf{m} \in \mathcal{M}} \sum_{\mathbf{x} \in \mathcal{X}} \mathbb{E}[\theta(\mathbf{M}_i, \mathbf{X}_i; \tilde{\psi}_C)\pi(\mathbf{M}_i, \mathbf{X}_i; \tilde{\psi}_S)\delta_i \mathbf{1}\{C_i = c\} S_i T_i \{Y_i - g_1(\mathbf{M}_i, \mathbf{X}_i; \psi_1^*)\} \mid \mathbf{M}_i = \mathbf{m}, \mathbf{X}_i = \mathbf{x}] \\
& \qquad \times \Pr(\mathbf{M}_i = \mathbf{m}, \mathbf{X}_i = \mathbf{x}) \\
= & \sum_{\mathbf{m} \in \mathcal{M}} \sum_{\mathbf{x} \in \mathcal{X}} \mathbb{E}[T_i \{Y_i - g_1(\mathbf{M}_i, \mathbf{X}_i; \psi_1^*)\} \mid S_i = 1, C_i = c, \mathbf{M}_i = \mathbf{m}, \mathbf{X}_i = \mathbf{x}] \\
& \qquad \times \delta_i \theta(\mathbf{M}_i, \mathbf{X}_i; \tilde{\psi}_C)\pi(\mathbf{M}_i, \mathbf{X}_i; \tilde{\psi}_S) \times \Pr(S_i = 1, C_i = c \mid \mathbf{M}_i = \mathbf{m}, \mathbf{X}_i = \mathbf{x}) \Pr(\mathbf{M}_i = \mathbf{m}, \mathbf{X}_i = \mathbf{x}) \\
= & \sum_{\mathbf{m} \in \mathcal{M}} \sum_{\mathbf{x} \in \mathcal{X}} \{\mathbb{E}[Y_i \mid T_i = 1, S_i = 1, C_i = c, \mathbf{M}_i = \mathbf{m}, \mathbf{X}_i = \mathbf{x}] - g_1(\mathbf{M}_i, \mathbf{X}_i; \psi_1^*)\} \\
& \qquad \times \theta(\mathbf{M}_i, \mathbf{X}_i; \tilde{\psi}_C)\pi(\mathbf{M}_i, \mathbf{X}_i; \tilde{\psi}_S) \times \Pr(S_i = 1, C_i = c \mid \mathbf{M}_i = \mathbf{m}, \mathbf{X}_i = \mathbf{x}) \Pr(\mathbf{M}_i = \mathbf{m}, \mathbf{X}_i = \mathbf{x}) \\
= & \ 0
\end{aligned}$$

which provides the desired result for the treatment group. We can use the similar proof for the expression for the control group. Therefore, this provides the proof for cases when the outcome models are correctly specified. This completes the proof. □

H.3 Proofs for Sign-Generalization

H.3.1 Proof: A Valid Test of the Union Null is valid for the Target Null.

We want to show that the under the target null $H_0^* : \mathbb{E}_{\mathcal{P}}\{Y_i^*(T=1, c) - Y_i^*(T=0, c)\} \leq 0$, $\Pr(\widetilde{p} \leq \alpha) \leq \alpha$ where $\widetilde{p} \equiv \max_k p_k$ and each p-value is valid for its corresponding null hypothesis.

First, under Assumption 5, the target null hypothesis implies the union of the K null hypotheses. Formally,

$$H_0^* \Rightarrow \bigcup_{k=1}^{K} H_0^k.$$

Under the union of the K null hypotheses, there is at least one true null hypothesis. We refer to it as H_0^ℓ and its corresponding p-value as p_ℓ. Then,

$$\begin{aligned} \Pr(\widetilde{p} \leq \alpha) &= \Pr\left\{ \bigcap_{k=1}^{K} (p_k \leq \alpha) \right\} \\ &\leq \Pr(p_\ell \leq \alpha) \leq \alpha. \end{aligned}$$

Taken together, under the target null hypothesis, $\Pr(\widetilde{p} \leq \alpha) \leq \alpha$, which completes the proof. □

H.3.2 Partial-Conjunction Test

In the partial conjunction test, we consider the following hypothesis.

$$\begin{aligned} & \widetilde{H}_0^r : \sum_{k=1}^{K} \mathbf{1}\{H_0^k \text{ is false}\} < r \\ \text{versus} \quad & \widetilde{H}_1^r : \sum_{k=1}^{K} \mathbf{1}\{H_0^k \text{ is false}\} \geq r \end{aligned} \tag{16}$$

For completeness, we reproduce all the necessary formal results on the partial conjunction test.

Result 1 (Validity of Partial Conjunction Test (Benjamini and Heller, 2008)) $\widetilde{p}_{(r)}$ is a valid p-value for the partial conjunction null hypothesis $\widetilde{H}_0^r$.

Proof. We want to show that under the partial conjunction null hypothesis $\widetilde{H}_0^r$, $\Pr(\widetilde{p}_{(r)} \leq \alpha) \leq \alpha$.

Under the partial conjunction null hypothesis, there are at least $K-r+1$ true null hypotheses. We use $q_1, \ldots, q_{K-r+1}$ to denote p-values corresponding to such true null hypotheses.

Now, we consider the main quantity.

$$\Pr(\widetilde{p}_{(r)} \leq \alpha) \leq \Pr\left\{(K-r+1)p_{(r)} \leq \alpha\right\} = \Pr\left(p_{(r)} \leq \frac{\alpha}{K-r+1}\right). \tag{17}$$

This implies that at least one of $\{q_1, \ldots, q_{K-r+1}\}$ is smaller than $\frac{\alpha}{K-r+1}$. Therefore,

$$\begin{aligned} \Pr(\widetilde{p}_{(r)} \leq \alpha) &\leq \Pr\left(p_{(r)} \leq \frac{\alpha}{K-r+1}\right) \\ &\leq \Pr\left\{ \bigcup_{i=1}^{K-r+1} \left(q_i \leq \frac{\alpha}{K-r+1}\right) \right\} \\ &\leq \sum_{i=1}^{K-r+1} \Pr\left(q_i \leq \frac{\alpha}{K-r+1}\right) \end{aligned}$$

$$\leq \quad \alpha,$$

where the first inequality comes from equation (17), the second from a definition of $\{q_1, \ldots, q_{K-r+1}\}$, the third from the union bound, and the final from the fact that each p-value is valid for its corresponding null hypothesis. This completes the proof. □

Result 2 (Reporting all thresholds (Benjamini and Heller, 2008))
No additional multiple testing correction is required when considering p-value $\widetilde{p}_{(r)}$ for all levels $r \in \{1, \ldots, K\}$. Formally, suppose the partial conjunction null holds when $r = s$, i.e., $\widetilde{H}_0^s$ is true. Then, $\Pr\{\bigcup_{r=s}^{K}(\widetilde{p}_{(r)} \leq \alpha)\} \leq \alpha$.

Proof. By the definition of $\widetilde{p}_{(r)}$, it satisfies the monotonicity, that is, $\widetilde{p}_{(r)} \leq \widetilde{p}_{(r+1)}$. Therefore, under the partial conjunction null $\widetilde{H}_0^s$,

$$\Pr\{\bigcup_{r=s}^{K}(\widetilde{p}_{(r)} \leq \alpha)\} = \Pr\{\widetilde{p}_{(s)} \leq \alpha\} \leq \alpha,$$

where the first equality follows from the monotonicity, and the second from the validity of the partial conjunction p-value (Result 1). □

Result 3 (Confidence Interval of True Non-Nulls (Benjamini and Heller, 2008))
Define r^* to be the number of true non-nulls. Then, $\Pr(r^* \in [r_{\max}, K]) \geq 1 - \alpha$ where $r_{\max} \equiv \max\{r : \widetilde{p}_{(r)} \leq \alpha\}$.

Proof. If $r^* = K$, then $\Pr(r^* \in [r_{\max}, K]) = 1$. Therefore, we consider cases where $r^* < K$.

$$\begin{aligned}
\Pr(r^* \in [r_{\max}, K]) &= \Pr(r^* \geq r_{\max}) \\
&= \Pr(\widetilde{p}_{(r^*+1)} > \alpha) \\
&= 1 - \Pr(\widetilde{p}_{(r^*+1)} \leq \alpha) \\
&\geq 1 - \alpha
\end{aligned}$$

where the first equality follows from the definition of the confidence interval, the second from the definition of $r_{\max}$, and the third from a rule of probability. When the true number of non-nulls is r^*, $\widetilde{H}_0^{r^*+1}$ holds. Therefore, $\Pr(\widetilde{p}_{(r^*+1)} \leq \alpha) \leq \alpha$, from which the final inequality follows. This completes the proof. □

I Simulations

To evaluate the performance of the T-PATE estimators, we conduct two sets of simulations. In our first set of simulations, we fully simulate the data generating process and control the parameterization of the sampling model and treatment effect heterogeneity. In our second set of simulations, we use the Broockman and Kalla (2016) data and CCES data as a basis for the sampling and treatment effect heterogeneity model. In combination, these simulations clarify conditions under which various estimators discussed in Section 5 can recover the T-PATE.

I.1 Full Simulations

I.1.1 Data Generating Process

Setup. In our first set of simulation, we fully control the data-generating process, including both the sampling and treatment effect heterogeneity models. We start by drawing a sample of size $N \in \{1000, 2000, 8000\}$. For each unit i, we draw ten pre-treatment covariates, $X_{i1}, \ldots, X_{i,10}$, each drawn independently from the standard normal distribution. We draw the experimental sample from our N original units with $S_i \sim \text{Bernoulli}(\pi_i)$ where S_i takes the value of one if unit i is sampled into the experimental sample and takes the value of zero otherwise. The sampling probability π is defined as

$$\Pr(S_i = 1 \mid \mathbf{X}_i) \equiv \pi_i = \frac{\exp(X_{i1} + \ldots + X_{i5})}{1 + \exp(X_{i1} + \ldots + X_{i5})}. \tag{18}$$

This sampling probability is 0.5 on average, implying that our experimental sample takes on sizes $n = \{500, 1000, 4000\}$, in expectation. We define our target sample as those units for which $S_i = 0$. Treatment T_i is assigned using complete randomization among the n experimental sample units.

We consider the two outcomes models: linear and non-linear outcome models.

Case 1: Linear Outcome Model Our first outcome model is linear in the pre-treatment covariates. In this model, we expect the outcome-based OLS estimator to perform well. We start by drawing coefficients $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_{10})$ where each element is independently drawn from the standard normal distribution.

We define the potential outcomes with the following system of linear equations. For each unit i,

$$\begin{aligned} Y_i(1) &= Y_i(0) + \tau_i \\ \tau_i &= X_{i1} + \ldots + X_{i5} + \epsilon_{i1} \\ Y_i(0) &= (1, \mathbf{X}_i)^\top \boldsymbol{\beta} + \epsilon_{i0} \end{aligned}$$

where we draw two error terms, ϵ_{i0} and ϵ_{i1}, from the independent standard normal for each unit. For each unit, we observe Y_i for the experimental sample as:

$$Y_i = T_i Y_i(1) + (1 - T_i) \cdot Y_i(0).$$

Case 2: Non-Linear Outcome Model Our second outcome model includes a non-linear relationship with the pre-treatment covariates, a scenario in which OLS should perform poorly, but the BART model can account for the non-linearities. The outcome model is based on the data-generating process considered in Hill (2011*a*). We start by drawing coefficients $\beta_1, \ldots, \beta_5$ randomly from $(0, 0.2, 0.4, 0.6, 0.5)$, each with equal probability, and $\beta_6, \ldots, \beta_{10}$ drawn independently from the standard normal. Let $\beta_0 = 0$. Finally, define an offset matrix, $\mathbf{W}$, with 5 columns and n rows with every value equal to 0.5.

We then define the potential outcomes with the following system of non-linear equations. For each unit i,

$$
\begin{aligned}
Y_i(1) &= (X_{i1}, \ldots, X_{i5})^\top \boldsymbol{\beta}_{[1:5]} + \epsilon_{i1} \\
Y_i(0) &= \exp((X_{i1}, \ldots, X_{i5})^\top \boldsymbol{\beta}_{[1:5]} + \mathbf{W}) + (X_{i6}, \ldots, X_{i10})^\top \boldsymbol{\beta}_{[6:10]} + \epsilon_{i0}
\end{aligned}
$$

where we draw two error terms, ϵ_{i0} and ϵ_{i1}, from the independent standard normal for each unit.

I.1.2 Estimators

We evaluate the three classes of estimators described in Section 5.1. In addition, we present the SATE estimator, using a difference-in-means within the experimental sample. For weighting-based estimators, we present the IPW and weighted least squares estimator. For outcome-based estimators, we use OLS and BART. Finally, for doubly-robust estimators, we use the augmented IPW estimator based on OLS and BART. For all estimators that incorporate outcome models, we use $(X_1, \ldots, X_5)$ to estimate outcome models.

We estimate sampling weights with logistic regression. We consider both scenarios in which the sampling model is either correctly or incorrectly specified. For the correctly specified sampling model, we use the correct set of variables $(X_1, \ldots, X_5)$. For the misspecified sampling model case, we use the incorrect set (X_1, X_2, X_3).

I.1.3 Results

Figure A8 presents results for 1000 simulations for each data generating process for the outcome model and for correct and incorrect sampling weights. Numerical summaries can be found in Table A1. When the sampling weights are correctly specified (purple), both the IPW and wLS estimators are consistent for the T-PATE regardless of the true outcome model. However, when sampling weights are misspecified (green), both weighting-based estimators exhibit a significant bias.

The performance of the outcome-based estimators depends on the underlying outcome model. The outcome-based estimator based on OLS consistently estimates the T-PATE when the true outcome model is linear. However, it exhibits significant bias when the true outcome model is non-linear. The outcome-based estimator based on BART performs well when the true outcome model is both linear and non-linear, although there is a small amount of residual finite sample bias in both cases.

Finally, the doubly robust estimators consistently estimate the T-PATE for both linear and non-linear outcomes when the sampling weights are correctly specified. For example, even though the outcome-based estimator based on OLS performs poorly when the true DGP is non-linear, the augmented IPW estimator with OLS is still consistent as far as the sampling model is correctly specified ("AIPW with OLS" in "Non-linear Outcome" with correct sampling weights; the first purple box plot in the fourth column in the bottom panel). This shows the benefit of the doubly robust estimators, which allow for consistent estimation even if one of the models (outcome or sampling) is misspecified. Even if the sampling weights are incorrect,

the doubly robust estimators perform well, if the outcome model is correctly specified. For example, the AIPW with OLS is consistent as far as the true outcome model is linear, even when sampling weights are incorrectly specified ("AIPW with OLS" in "Linear Outcome" with wrong sampling weights; the first green box plot in the fourth column in the top panel). Similarly, the AIPW with BART is consistent if the outcome model is correctly specified (non-linear), or the sampling weights are correctly specified.

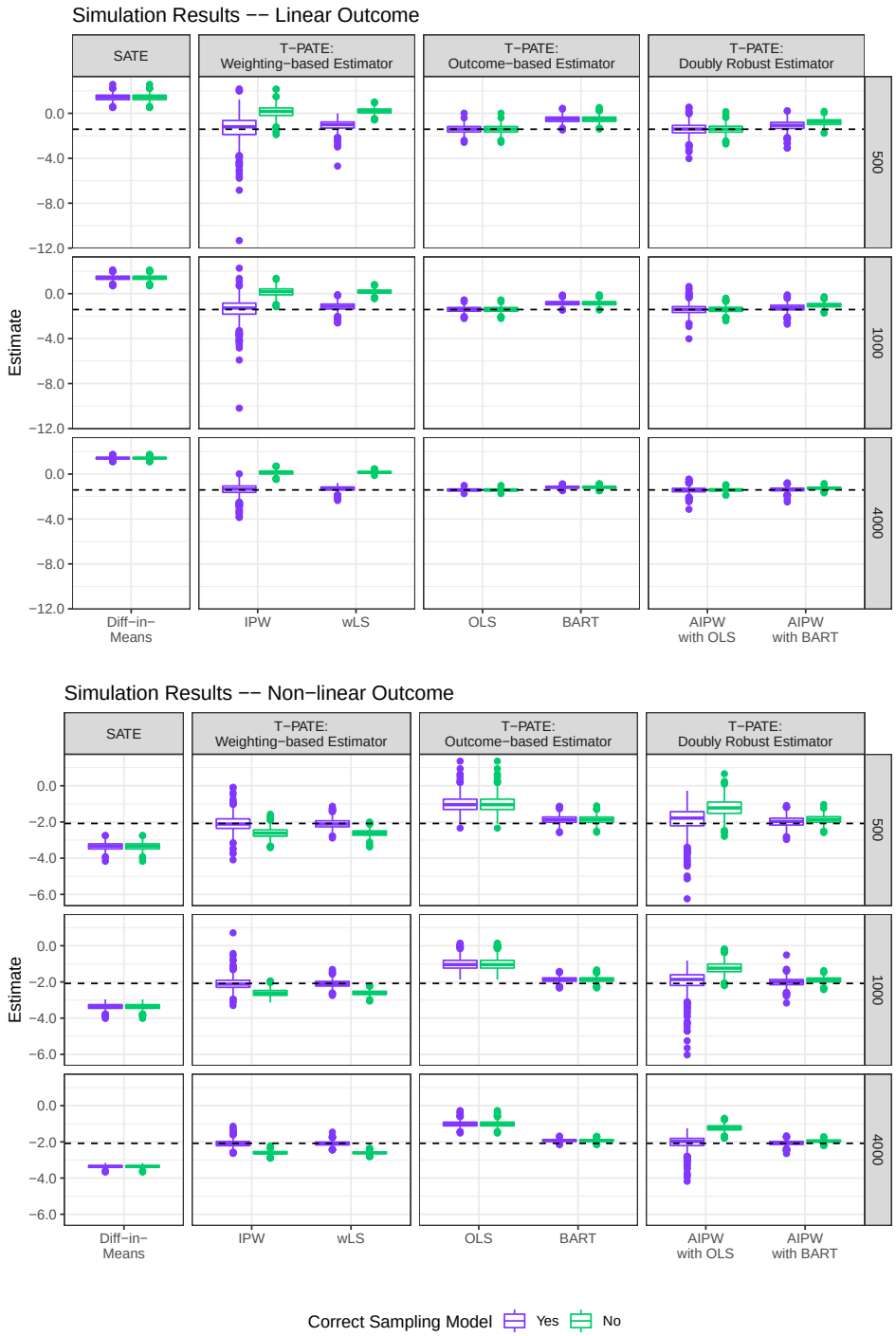

Figure A8: Full Simulation Results. The top panel presents results when the true outcome model is linear; the bottom panel presents results when the true outcome model is non-linear. Results for the correct sampling model are presented in purple, and those for the incorrectly specified sampling model are presented in green. In each figure, the true T-PATE is denoted by the horizontal dashed line. Rows in each panel denotes sample size of the experimental data, $n = \{500, 1000, 4000\}$.

Table A1: Numeric Values for Simulations in Figure A8.

		Linear Outcome				Non-Linear Outcome			
Estimator	n	Bias	SE	MSE	Avg. Interval Length	Bias	SE	MSE	Avg. Interval Length
Correct Sampling Model									
	500	2.834	0.289	8.114	1.148	-1.266	0.206	1.644	0.812
Diff-in-Means	1000	2.831	0.207	8.058	0.809	-1.275	0.152	1.650	0.577
	4000	2.823	0.103	7.981	0.405	-1.275	0.074	1.631	0.290
	500	0.113	1.099	1.219	3.446	-0.008	0.421	0.177	1.352
IPW	1000	0.035	0.888	0.789	2.831	-0.009	0.327	0.107	1.072
	4000	0.034	0.493	0.244	1.660	-0.005	0.191	0.037	0.639
	500	0.364	0.442	0.328	1.466	-0.019	0.252	0.064	0.871
wLS	1000	0.256	0.340	0.181	1.190	-0.010	0.198	0.039	0.681
	4000	0.104	0.219	0.059	0.779	-0.006	0.119	0.014	0.413
	500	-0.004	0.355	0.126	0.000	1.074	0.446	1.351	0.000
OLS	1000	0.014	0.239	0.057	0.000	1.070	0.322	1.247	0.000
	4000	-0.003	0.122	0.015	0.000	1.072	0.162	1.175	0.000
	500	0.886	0.274	0.861	0.801	0.213	0.205	0.088	0.591
BART	1000	0.567	0.202	0.363	0.574	0.212	0.143	0.066	0.433
	4000	0.242	0.100	0.069	0.294	0.158	0.070	0.030	0.219
	500	0.006	0.540	0.292	0.000	0.208	0.675	0.498	0.000
AIPW with OLS	1000	0.009	0.425	0.181	0.000	0.137	0.554	0.326	0.000
	4000	-0.014	0.254	0.064	0.000	0.045	0.349	0.124	0.000
	500	0.330	0.423	0.288	1.314	0.099	0.279	0.088	0.946
AIPW with BART	1000	0.170	0.321	0.132	1.013	0.069	0.224	0.055	0.727
	4000	0.035	0.182	0.034	0.540	0.022	0.126	0.016	0.391
Incorrect Sampling Model									
	500	2.834	0.289	8.114	1.148	-1.266	0.206	1.644	0.812
Diff-in-Means	1000	2.831	0.207	8.058	0.809	-1.275	0.152	1.650	0.577
	4000	2.823	0.103	7.981	0.405	-1.275	0.074	1.631	0.290
	500	1.561	0.524	2.710	2.055	-0.523	0.259	0.341	0.983
IPW	1000	1.573	0.389	2.624	1.465	-0.522	0.186	0.307	0.703
	4000	1.550	0.194	2.440	0.751	-0.520	0.096	0.280	0.361
	500	1.624	0.253	2.702	0.986	-0.527	0.177	0.309	0.692
wLS	1000	1.600	0.188	2.594	0.726	-0.525	0.134	0.293	0.506
	4000	1.563	0.097	2.451	0.386	-0.520	0.071	0.276	0.264
	500	-0.004	0.355	0.126	0.000	1.074	0.446	1.351	0.000
OLS	1000	0.014	0.239	0.057	0.000	1.070	0.322	1.247	0.000
	4000	-0.003	0.122	0.015	0.000	1.072	0.162	1.175	0.000
	500	0.888	0.275	0.864	0.805	0.213	0.203	0.086	0.593
BART	1000	0.566	0.203	0.362	0.578	0.214	0.146	0.067	0.429
	4000	0.242	0.102	0.069	0.296	0.157	0.069	0.030	0.221
	500	0.005	0.391	0.153	0.000	0.878	0.446	0.970	0.000
AIPW with OLS	1000	0.013	0.271	0.073	0.000	0.860	0.331	0.849	0.000
	4000	-0.003	0.137	0.019	0.000	0.854	0.168	0.757	0.000
	500	0.634	0.296	0.489	0.926	0.223	0.223	0.100	0.652
AIPW with BART	1000	0.385	0.214	0.194	0.647	0.191	0.156	0.061	0.477
	4000	0.155	0.110	0.036	0.326	0.122	0.075	0.020	0.243

I.2 Naturalistic Simulations

While the analyses in Section I.1 clarify conditions under which the three classes of estimators are consistent for the T-PATE, we now turn to more naturalistic simulations to better evaluate their performance in social science data. We build our simulations on the Broockman and Kalla (2016) experimental sample and the CCES data for Florida.

I.2.1 Data Generating Process

As with our full simulations above, we consider two scenarios for the outcome model, a linear and non-linear case. For each unit i, we define a vector of covariates, $\mathbf{X}_i$, using gender, race, age (in years), ideology, party identification, and religiosity. We use these pre-treatment covariates in the estimation of the treatment effect heterogeneity model and sampling model.

Case 1: Linear Outcome Model For our linear outcome model case, we use OLS to estimate treatment effect heterogeneity in the experimental sample of Broockman and Kalla (2016). In particular, we construct our linear outcome model by the following steps:

1. Estimate treatment effect heterogeneity within the experimental data using OLS separately for the treated and control group using the experimental data from Broockman and Kalla (2016).

2. Use the predictions from the estimated model defined in the first step to construct the potential outcome under control $Y_i(0)$ and the individual level treatment effect τ_i on the target population defined by the CCES.

3. Rescale τ_i to have mean 1 in the target population data.

4. Construct $Y_i(1) = Y_i(0) + \tau_i$.

5. Re-estimate treatment effect heterogeneity within the experimental data using OLS on the adjusted $Y_i(0)$ and $Y_i(1)$ from above and, construct $Y_i(1)$ and $Y_i(0)$ from the re-estimated model.

Case 2: Non-Linear Outcome Model For the non-linear outcome model, we use BART to flexibly estimate treatment effect heterogeneity within the experiment. We construct our non-linear outcome model by the following steps.

1. Estimate treatment effect heterogeneity within the experimental data from Broockman and Kalla (2016) using `bartc` function, from the `bartCause` package, with default parameters.

2. Use the predictions from the estimated model defined in the first step to construct the potential outcome under control $Y_i(0)$ and the individual level treatment effect τ_i on the target population defined by the CCES.

3. Rescale τ_i to have mean 1 in the target population data.

4. Construct $Y_i(1) = Y_i(0) + \tau_i$.

5. Re-estimate treatment effect heterogeneity within the experimental data using BART on the adjusted $Y_i(0)$ and $Y_i(1)$ from above and, construct $Y_i(1)$ and $Y_i(0)$ from the re-estimated model.

As discussed in the original manuscript, there is limited treatment effect heterogeneity within the original experimental data. In order to induce bias in our experimental sample, we want to make sure there is a strong correlation between the sampling probability π_i and the estimated individual level treatment effect. To do this, rather than model the difference between the true experimental sample and the CCES, we construct a pseudo-experimental sample based on the treatment effect size. The probability of being included in this sample depends on the outcome model.

For the true linear outcome model, we take one draw from the CCES to construct an experimental sample where units are included with probability 0.035 if they are in the bottom 75% of treatment effects and 0.01 if they are in the top 25%. S_i denotes inclusion in this sample. We then model the sampling probability π_i using logit of the indicator S_i on $\mathbf{X}_i$. This sampling probability is used to draw an experimental sample from the CCES within each simulation.

BART estimates much less treatment effect heterogeneity than OLS. In order to scale the bias of the SATE to be similar across the two models, we update the probabilities we use when constructing S_i. Units are included with probability 0.05 if they are in the bottom 95% of treatment effects and 0.95 if they are in the top 5%. S_i denotes inclusion in this sample. We then construct π_i using logit of the indicator S_i on $\mathbf{X}_i$.

Finally, within each simulation, we draw a random sample of size 5000 from the CCES data that serves as our target population. As the experimental data, we draw a fixed sample of size $n = \{500, 1000, 4000\}$ using π_i defined as above. Within each simulation, potential outcomes are constructed using $Y_i(1)$ and $Y_i(0)$ as defined above for each outcome model, plus random noise. Treatment T_i is assigned using complete randomization among the n experimental sample units.

I.2.2 Estimators

We evaluate the three classes of estimators described in Section 5.1. In addition, we present the SATE estimator, using a difference-in-means within the experimental sample. For weighting-based estimators, we present the IPW and weighted least squares estimator. For outcome-based estimators, we use OLS and BART. Finally, for doubly-robust estimators, we use the augmented IPW estimator based on OLS and BART. For all estimators that incorporate outcome models, we use $\mathbf{X}_i$ to estimate outcome models.

We estimate sampling weights with logistic regression. We consider both scenarios in which the sampling model is either correctly or incorrectly specified. For the correctly specified sampling model, we use the correct set of variables–all variables in $\mathbf{X}_i$. For the misspecified sam-

pling model case, we only use religiosity and party identification to estimate logistic regression and construct sampling weights.

I.2.3 Results

Figure A9 presents results for 1000 simulations from the linear and non-linear outcome model data-generating processes. Numerical summaries can be found in Table A2. In this naturalistic simulation, we confirm the same pattern we show theoretically and in the previous full simulation. Weighting-based estimators (the second column in both panels) are consistent when sampling weights are correctly specified (purple), but they are not consistent when weights are misspecified (green).

In this naturalistic simulation, because there is limited treatment effect heterogeneity with respect to $\mathbf{X}$, the outcome-based estimator with OLS performs well even when the outcome is non-linear, implying the treatment effect heterogeneity induced by BART is well approximated by a fully-interacted linear model ("OLS" estimator in the third column of both panels). As with the full simulations, we see some finite sample bias, decreasing with sample size, using the BART estimator in both linear and non-linear outcome data generating processes.

Results confirm that the doubly robust estimators are consistent even when one of the two models — outcome and sampling models — is misspecified.

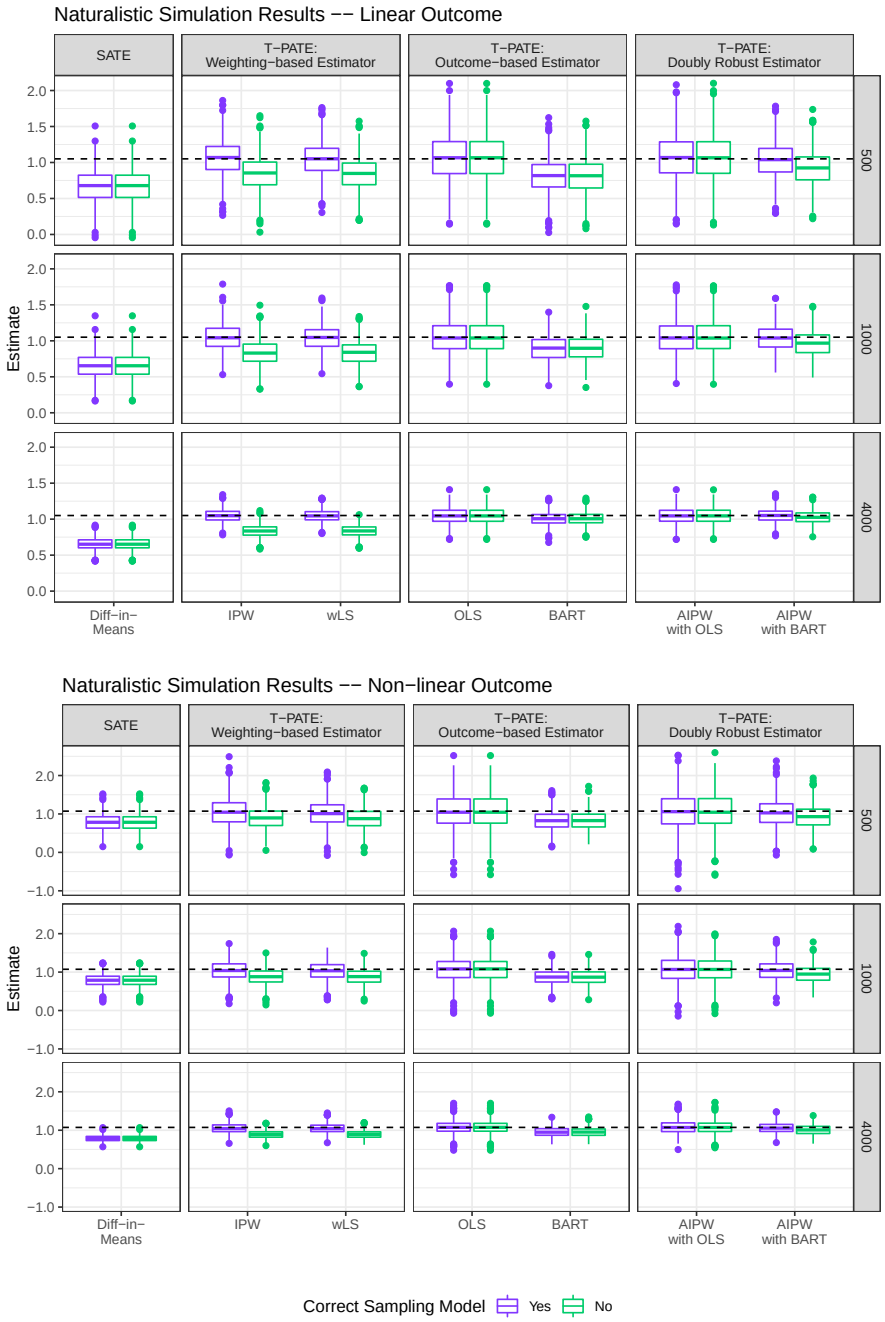

Figure A9: Naturalistic Simulation Results. In the top panel, we report results when the true outcome model is linear; in the bottom panel, we present results when the true outcome model is non-linear. Purple represents results when the sampling weights are correctly specified, and green represents results when the sampling weights are incorrectly specified. In each figure, the true T-PATE is denoted by the horizontal dashed line. In each panel, rows indicate the experimental sample sizes $n = \{500, 1000, 4000\}$.

Table A2: Numeric Values for Simulations in Figure A9.

		Linear Outcome				Non-Linear Outcome			
Estimator	n	Bias	SE	MSE	Avg. Interval Length	Bias	SE	MSE	Avg. Interval Length
Correct Sampling Model									
	500	-0.383	0.233	0.201	0.941	-0.290	0.218	0.131	0.872
Diff-in-Means	1000	-0.399	0.172	0.189	0.664	-0.287	0.158	0.107	0.616
	4000	-0.395	0.084	0.163	0.332	-0.286	0.077	0.088	0.308
	500	0.009	0.246	0.060	1.016	-0.026	0.357	0.128	1.354
IPW	1000	-0.006	0.178	0.032	0.713	-0.036	0.250	0.064	0.961
	4000	-0.002	0.089	0.008	0.354	-0.022	0.127	0.017	0.496
	500	-0.009	0.230	0.053	0.965	-0.054	0.328	0.111	1.242
wLS	1000	-0.011	0.171	0.029	0.673	-0.046	0.241	0.060	0.903
	4000	-0.003	0.084	0.007	0.332	-0.025	0.123	0.016	0.477
	500	0.012	0.334	0.112	0.000	-0.001	0.457	0.209	0.000
OLS	1000	0.001	0.232	0.054	0.000	-0.001	0.328	0.107	0.000
	4000	-0.002	0.114	0.013	0.000	0.002	0.159	0.025	0.000
	500	-0.239	0.240	0.115	0.662	-0.247	0.243	0.120	0.710
BART	1000	-0.155	0.178	0.056	0.481	-0.203	0.191	0.077	0.543
	4000	-0.045	0.089	0.010	0.266	-0.117	0.124	0.029	0.336
	500	0.011	0.335	0.112	0.000	0.006	0.492	0.242	0.000
AIPW with OLS	1000	0.000	0.232	0.054	0.000	-0.003	0.353	0.124	0.000
	4000	-0.002	0.114	0.013	0.000	0.004	0.174	0.030	0.000
	500	-0.017	0.243	0.059	0.691	-0.044	0.357	0.129	0.990
AIPW with BART	1000	-0.014	0.180	0.033	0.490	-0.035	0.256	0.067	0.700
	4000	-0.001	0.090	0.008	0.271	-0.013	0.133	0.018	0.377
Incorrect Sampling Model									
	500	-0.383	0.233	0.201	0.941	-0.290	0.218	0.131	0.872
Diff-in-Means	1000	-0.399	0.172	0.189	0.664	-0.287	0.158	0.107	0.616
	4000	-0.395	0.084	0.163	0.332	-0.286	0.077	0.088	0.308
	500	-0.201	0.242	0.099	0.980	-0.180	0.287	0.114	1.114
IPW	1000	-0.217	0.176	0.078	0.690	-0.190	0.202	0.077	0.781
	4000	-0.213	0.086	0.053	0.343	-0.182	0.103	0.044	0.394
	500	-0.209	0.229	0.096	0.933	-0.192	0.273	0.112	1.074
wLS	1000	-0.219	0.169	0.076	0.652	-0.194	0.198	0.077	0.757
	4000	-0.214	0.081	0.052	0.322	-0.183	0.100	0.044	0.383
	500	0.012	0.334	0.112	0.000	-0.001	0.457	0.209	0.000
OLS	1000	0.001	0.232	0.054	0.000	-0.001	0.328	0.107	0.000
	4000	-0.002	0.114	0.013	0.000	0.002	0.159	0.025	0.000
	500	-0.241	0.238	0.115	0.682	-0.244	0.242	0.118	0.720
BART	1000	-0.150	0.174	0.053	0.487	-0.204	0.192	0.078	0.552
	4000	-0.043	0.089	0.010	0.266	-0.117	0.123	0.029	0.338
	500	0.012	0.334	0.112	0.000	-0.001	0.464	0.215	0.000
AIPW with OLS	1000	0.001	0.232	0.054	0.000	-0.005	0.333	0.111	0.000
	4000	-0.002	0.114	0.013	0.000	0.002	0.163	0.027	0.000
	500	-0.132	0.242	0.076	0.683	-0.150	0.298	0.111	0.858
AIPW with BART	1000	-0.092	0.181	0.041	0.488	-0.126	0.228	0.068	0.622
	4000	-0.025	0.090	0.009	0.270	-0.063	0.130	0.021	0.352

J Literature Review of *American Political Science Review*

To evaluate current practice for addressing concerns about external validity, we conducted a review of the five most recent years of all articles that use randomized experiments and common observational causal designs that are published in the *American Political Science Review* (APSR).

To conduct our review for experiments, using the advanced search on the APSR website, we searched for all articles that mention "experiment" in the years 2015-2019 (inclusive). We read each article to determine if the author(s) used a randomized experiment. This resulted in 35 articles, outlined in Table A3. We then coded each article for the type of experiment, and found 18 field, 3 lab, and 14 survey experiments.

For observational studies, we review papers that use instumental variables, the regression discontinuity design, or the difference-in-differences design. To find papers using instrumental variables, we searched for all articles that mention "instumental variable". To find papers using the regression discontinuity design, we searched for all articles that mention "regression discontinuity". To find papers using difference-in-differences, we searched for all articles that mention "difference-in-difference" and "two-way fixed effect." We read each article to determine if the author(s) used an appropriate observational causal design. This resulted in 20 articles which use instrumental variables, 16 which use regression discontinuity designs and 10 that use differences-in-differences (we focus on papers that uses the basic DID design and the staggered adoption design). These references are outlined in Table A3.

With regards to external validity, we reviewed two dimensions: (1) formal analysis of external validity and (2) use of purposive variations. There were 4 experimental articles (11%) that conducted some formal analysis in the main text to address external validity. All of these papers were survey experiments and used survey weights to adjust for sample representativeness for X-validity. As for observational studies, there were 6 papers out of 46 papers (13%) conducted some formal analysis in the main text to address external validity. In particular, there was 1 instrumental variables article (5%), 3 regression discontinuity design articles (19%), and 2 difference-in-differences articles (10%). Most of these involved running the analysis on a larger data set that included more contextual variation. While we found very few formal analyses, we do note that most authors, across methods, do informally discuss the implications of their findings for external validity.

There were 29 experimental articles (83%) that included some purposive variations in one of four dimensions of external validity. There were 41 observational studies that included some purposive variations (89%). In particular, there were 16 instrumental variables articles (80%). 15 regression discontinuity design articles (94%), and 10 differences-in-differences articles (100%).

Table A3: Randomized experiments and observational studies in the *APSR* from 2015-2019.

Authors	Year	Title
Experiments		
Allison P. Anoll	2018	What Makes a Good Neighbor? Race, Place, and Norms of Political Participation
Eric Arias and Pablo Balán and Horacio Larreguy and John Marshall and Pablo Querubin	2019	Information Provision, Voter Coordination, and Electoral Accountability: Evidence from Mexican Social Networks
Adam Michael Auerbach and Tariq Thachil	2018	How Clients Select Brokers: Competition and Choice in India's Slums
Alexandra Avdeenko and Michael J. Gilligan	2015	International Interventions to Build Social Capital: Evidence from a Field Experiment in Sudan
Andy Baker	2015	Race, Paternalism, and Foreign Aid: Evidence from U.S. Public Opinion
Robert A. Blair and Sabrina M. Karim and Benjamin S. Morse	2019	Establishing the Rule of Law in Weak and War-torn States: Evidence from a Field Experiment with the Liberian National Police
Christopher Blattman and Jeannie Annan	2016	Can Employment Reduce Lawlessness and Rebellion? A Field Experiment with High-Risk Men in a Fragile State
Pazit ben-nun Bloom and Gizem Arikan and Marie Courtemanche	2015	Religious Social Identity, Religious Belief, and Anti-Immigration Sentiment
Céline Braconnier and Jean-yves Dormagen and Vincent Pons	2017	Voter Registration Costs and Disenfranchisement: Experimental Evidence from France
Daniel M. Butler and Hans J.g. Hassell	2018	On the Limits of Officials' Ability to Change Citizens' Priorities: A Field Experiment in Local Politics
Taylor N. Carlson	2019	Through the Grapevine: Informational Consequences of Interpersonal Political Communication
Nicholas Carnes and Noam Lupu	2016	Do Voters Dislike Working-Class Candidates? Voter Biases and the Descriptive Underrepresentation of the Working Class

Table A3: Randomized experiments and observational studies in the *APSR* from 2015-2019.

Authors	**Year**	**Title**
James D. Fearon and Macartan Humphreys and Jeremy M. Weinstein	2015	How Does Development Assistance Affect Collective Action Capacity? Results from a Field Experiment in Post-Conflict Liberia
Jens Grosser and Thomas R. Palfrey	2019	Candidate Entry and Political Polarization: An Experimental Study
Guy Grossman and Kristin Michelitch	2018	Information Dissemination, Competitive Pressure, and Politician Performance between Elections: A Field Experiment in Uganda
Hahrie Han	2016	The Organizational Roots of Political Activism: Field Experiments on Creating a Relational Context
Andrew Healy and Katrina Kosec and Cecilia Hyunjung Mo	2017	Economic Development, Mobility, and Political Discontent: An Experimental Test of Tocqueville's Thesis in Pakistan
Alexander Hertel-fernandez and Matto Mildenberger and Leah C. Stokes	2019	Legislative Staff and Representation in Congress
John B. Holbein	2017	Childhood Skill Development and Adult Political Participation
Leonie Huddy and Lilliana Mason and Lene Aaroe	2015	Expressive Partisanship: Campaign Involvement, Political Emotion, and Partisan Identity
Joshua L. Kalla and David E. Broockman	2018	The Minimal Persuasive Effects of Campaign Contact in General Elections: Evidence from 49 Field Experiments
Amy E. Lerman and Meredith L. Sadin and Samuel Trachtman	2017	Policy Uptake as Political Behavior: Evidence from the Affordable Care Act
Zhao Li	2018	How Internal Constraints Shape Interest Group Activities: Evidence from Access-Seeking PACs
Edmund Malesky and Markus Taussig	2019	Participation, Government Legitimacy, and Regulatory Compliance in Emerging Economies: A Firm-Level Field Experiment in Vietnam

Table A3: Randomized experiments and observational studies in the *APSR* from 2015-2019.

Authors	**Year**	**Title**
Neil Malhotra and Benoît Monin and Michael Tomz	2019	Does Private Regulation Preempt Public Regulation?
Kristin Michelitch	2015	Does Electoral Competition Exacerbate Interethnic or Interpartisan Economic Discrimination? Evidence from a Field Experiment in Market Price Bargaining
Alexandra Scacco and Shana S. Warren	2018	Can Social Contact Reduce Prejudice and Discrimination? Evidence from a Field Experiment in Nigeria
Gabor Simonovits and Gabor Kezdi and Peter Kardos	2018	Seeing the World Through the Other's Eye: An Online Intervention Reducing Ethnic Prejudice
Dawn Langan Teele and Joshua Kalla and Frances Rosenbluth	2018	The Ties That Double Bind: Social Roles and Women's Underrepresentation in Politics
Ali A. Valenzuela and Melissa R. Michelson	2016	Turnout, Status, and Identity: Mobilizing Latinos to Vote with Group Appeals
Dalston G. Ward	2019	Public Attitudes toward Young Immigrant Men
Ariel R. White and Noah L. Nathan and Julie K. Faller	2015	What Do I Need to Vote? Bureaucratic Discretion and Discrimination by Local Election Officials
Jonathan Woon	2018	Primaries and Candidate Polarization: Behavioral Theory and Experimental Evidence
Lauren E. Young	2019	The Psychology of State Repression: Fear and Dissent Decisions in Zimbabwe
Adam Zelizer	2019	Is Position-Taking Contagious? Evidence of Cue-Taking from Two Field Experiments in a State Legislature
Regression-Discontinuity-Designs		
Jo Dahlgaard	2018	Trickle-Up Political Socialization: The Causal Effect on Turnout of Parenting a Newly Enfranchised Voter.

Table A3: Randomized experiments and observational studies in the *APSR* from 2015-2019.

Authors	**Year**	**Title**
Jon H. Fiva and Daniel M. Smith	2018	Political Dynasties and the Incumbency Advantage in Party-Centered Environments.
Olle Folke and Torsten Persson and Johanna Rickne	2016	The Primary Effect: Preference Votes and Political Promotions.
Jacob M. Grumbach and Alexander Sahn	2020	Race and Representation in Campaign Finance.
Saad Gulzar and Benjamin J. Pasquale	2017	Politicians, Bureaucrats, and Development: Evidence from India.
Jens Hainmueller and Dominik Hangartner and Giuseppe Pietrantuono	2017	Catalyst or Crown: Does Naturalization Promote the Long-Term Social Integration of Immigrants?
Andrew B. Hall and Daniel M. Thompson	2018	Who Punishes Extremist Nominees? Candidate Ideology and Turning Out the Base in US Elections.
Andrew B. Hall	2015	What Happens When Extremists Win Primaries?
John Holbein	2016	Left Behind? Citizen Responsiveness to Government Performance Information.
Instrumental Variables		
Robert Braun	2016	Religious Minorities and Resistance to Genocide: The Collective Rescue of Jews in the Netherlands during the Holocaust.
Lars-Erik Cederman and Simon Hug and andreas Schädel and Julian Wucherpfennig	2015	Territorial Autonomy in the Shadow of Conflict: Too Little, Too Late?
Italo Colantone and Piero Stanig	2018	Global Competition and Brexit.
Kevin Croke and Guy Grossman and Horacio A. Larreguy and John Marshall	2016	Deliberate Disengagement: How Education Can Decrease Political Participation in Electoral Authoritarian Regimes.
Aditya Dasgupta	2018	Technological Change and Political Turnover: The Democratizing Effects of the Green Revolution in India.
Michael T. Dorsch and Paul Maarek	2019	Democratization and the Conditional Dynamics of Income Distribution.

Table A3: Randomized experiments and observational studies in the *APSR* from 2015-2019.

Authors	**Year**	**Title**
Paul Castañeda Dower and Evgeny Finkel and Scott Gehlbach and Steven Nafziger	2018	Collective Action and Representation in Autocracies: Evidence from Russia's Great Reforms.
David Doyle	2015	Remittances and Social Spending.
Anselm Hager and Krzysztof Krakowski and Max Schaub	2019	Ethnic Riots and Prosocial Behavior: Evidence from Kyrgyzstan.
Jens Hainmueller and Dominik Hangartner and Giuseppe Pietrantuono	2017	Catalyst or Crown: Does Naturalization Promote the Long-Term Social Integration of Immigrants?
Dominik Hangartner and Elias Dinas and Moritz Marbach and Konstantinos Matakos and Dimitrios Xefteris	2019	Does Exposure to the Refugee Crisis Make Natives More Hostile?
Ari Hyytinen and Jaakko Meriläinen and Tuukka Saarimaa and Otto Toivanen and Janne Tukiainen	2018	Public Employees as Politicians: Evidence from Close Elections.
Sacha Kapoor and Arvind Magesan	2018	Independent Candidates and Political Representation in India.
David D. Laitin and Rajesh Ramachandran	2016	Language Policy and Human Development.
Gareth Nellis and Niloufer Siddiqui	2018	Secular Party Rule and Religious Violence in Pakistan.
Emily Hencken Ritter and Courtenay R. Conrad	2016	Preventing and Responding to Dissent: The Observational Challenges of Explaining Strategic Repression.
Ariel White	2019	Misdemeanor Disenfranchisement? The Demobilizing Effects of Brief Jail Spells on Potential Voters.
Lucas Leemann and Fabio Wasserfallen	2016	The Democratic Effect of Direct Democracy.
Karl-Oskar Lindgren and Sven Oskarsson and Mikael Persson	2019	Enhancing Electoral Equality: Can Education Compensate for Family Background Differences in Voting Participation?
Arturas Rozenas and Yuri M. Zhukov	2019	Mass Repression and Political Loyalty: Evidence from Stalin's 'Terror by Hunger'.

Table A3: Randomized experiments and observational studies in the *APSR* from 2015-2019.

Authors	Year	Title
Differences-in-differences		
Diana Z. O'Brien and Johanna Rickne	2016	Gender Quotas and Women's Political Leadership
Francisco Garfias	2018	Elite Competition and State Capacity Development: Theory and Evidence from Post-Revolutionary Mexico
Gregory J. Martin and Joshua Mccrain	2019	Local News and National Politics
Jens Blom-Hansen and Kurt Houlberg and Søren Serritzlew and Daniel Treisman	2016	Jurisdiction Size and Local Government Policy Expenditure: Assessing the Effect of Municipal Amalgamation
Joshua D. Clinton and Michael W. Sances	2018	The Politics of Policy: The Initial Mass Political Effects of Medicaid Expansion in the States
Martin Vinæs Larsen and Frederik Hjorth and Peter Thisted Dinesen and Kim Mannemar Sønderskov	2019	When Do Citizens Respond Politically to the Local Economy? Evidence from Registry Data on Local Housing Markets
Michael Becher and Irene Menéndez González	2019	Electoral Reform and Trade-Offs in Representation
Peter Selb and Simon Munzert	2018	Examining a Most Likely Case for Strong Campaign Effects
Ryan D. Enos and Aaron R. Kaufman and Melissa L. Sands	2019	Can Violent Protest Change Local Policy Support?
Vasiliki Fouka	2019	How Do Immigrants Respond to Discrimination?

K Numeric Results and Model Specification

In this section we provide numerical results for all figures containing analytical results in the main manuscript and appendices. Where appropriate, we also provide the associated model specification. For ease of reference, we include a brief description of the analysis and results and the associated figure reference.

K.1 Results for Broockman and Kalla (2016) analysis in Figure 7

Figure 7 presents point estimates and their 95% confidence intervals using different T-PATE estimators. The codebook and original authors' replication files can be found at https://doi.org/10.7910/DVN/WK
The associated models used in our analyses are described below.

- SATE: Following the original authors, we estimate the SATE for the trans-tolerance index at time t with regression controls pre-specified in the authors' pre-analysis plan and replication code as:

$$\begin{aligned}\text{transtolerance}_{ti} \sim\ & T_i + \text{miami_trans_law_t0} + \text{miami_trans_law2_t0} + \text{therm_trans_t0} \\ & + \text{gender_norms_sexchange_t0} + \text{gender_norms_moral_t0} \\ & + \text{gender_norms_abnormal_t0} + \text{ssm_t0} + \text{therm_obama_t0} \\ & + \text{therm_gay_t0} + \text{vf_democrat} + \text{ideology_t0} \\ & + \text{religious_t0} + \text{exposure_gay_t0} + \text{exposure_trans_t0} \\ & + \text{pid_t0} + \text{sdo_scale} + \text{gender_norm_daugher_t0} \\ & + \text{gender_norm_looks_t0} + \text{gender_norm_rights_t0} + \text{therm_afams_t0} \\ & + \text{vf_female} + \text{vf_hispanic} + \text{vf_black} + \text{vf_age} \\ & + \text{survey_language_es} + \text{cluster_level_t0_scale_mean}\end{aligned}$$

  where T_i is the treatment indicator. All linear regressions are estimated using `lm_robust` (Blair et al., 2019), with bootstrapped standard errors.

- IPW: Our IPW estimator uses calibration weights in which we calibrate on the following variables:

$$\mathbf{X} = [\text{vf_female}, \text{vf_black}, \text{vf_white}, \text{religious_t0}, \text{pid_t0}, \text{vf_age_bucket}]$$

  Let P_x be defined as the vector of population means for each variable (with categorical variables coded as indicators for each level).

  We construct calibration (or balancing) weights such that

$$\begin{aligned}\min_{w} \quad & \sum w_i log(w_i) \\ \text{subject to} \quad & \sum w_i X_i = P_x, \\ & \sum w_i = 1, \text{ and } \ 0 \leq w_i \leq 1.\end{aligned}$$

  See Deville and Särndal (1992), Hainmueller (2012), or Hartman et al. (2015) for more details about calibration weighting. We then conduct weighted least squares of $\text{transtolerance}_{ti} \sim T_i$, with bootstrapped standard errors. The model is implemented using our associated package with the function `tpate` with settings `est_type = "ipw"` and `weights_type = "calibration"`. The underlying calibration code relies on the `calibrate` function in the `survey` package (Lumley, 2020) with default settings. See Section 5.1.1.

- wLS: The weighted least squares analysis builds on the IPW model, running the same regression additionally with the inclusion of the regression controls included in the SATE

estimator, which were pre-specified by the original authors. We calculate bootstrapped standard errors. The model is implemented using our associated package with the function `tpate` with settings `est_type = "wls"` and `weights_type = "calibration"`. See Section 5.1.1.

- OLS: The OLS outcome-based estimator estimates uses the following specification separately for the treated and control groups:

$$\begin{aligned}\text{transtolerance}_{ti} \sim\ & \text{vf_age} + \text{vf_female} + \text{vf_black} \\ & + \text{vf_white} + \text{religious_t0} + \text{ideology_t0} + \text{pid_t0}\end{aligned}$$

  The resulting coefficients are used to project and calculate as the average predicted outcome under treatment and control (respectively) using the covariate distribution of the population. The effect is estimated as the average difference in these average predicted outcomes. We use bootstrapped standard errors. The model is implemented using our associated package with the function `tpate` with settings `est_type = "outcome-ols"`. See Section 5.1.2.

- BART: For the BART outcome-based estimator, we estimate the model:

$$Y = f(t, x) + \epsilon \qquad \epsilon \sim N(0, \sigma^2)$$

  where t is the treatment indicator and the x are the regression controls included in the OLS outcome-based estimator.

  The model is implemented using our associated package with the function `tpate` with settings `est_type = "outcome-bart"`. The underlying code calls `bartc` in the `bartCause` package (Hill, 2011*b*) with default settings. Credible intervals are calculated over the posterior. See Section 5.1.2.

- AIPW with OLS: The augmented OLS estimator uses the regression specification outlined under "OLS" and the calibration weights described in "IPW". We use bootstrapped standard errors. The model is implemented using our associated package with the function `tpate` with settings `est_type = "dr-ols"` and `weights_type = "calibration"`. See Section 5.1.3.

- AIPW with BART: The augmented BART estimator uses the specification outlined under "BART" and the calibration weights described in "IPW". Credible intervals are calculated over the posterior. The model is implemented using our associated package with the function `tpate` with settings `est_type = "dr-bart"` and `weights_type = "calibration"`. See Section 5.1.3.

The numeric results for Figure 7 are presented in Table A4 below.

Table A4: Numeric Values for T-PATE Estimates for Broockman and Kalla (2016).

Estimator Type	Estimator	Estimate	SE	95% CI
Measurement at +3 Days				
SATE	OLS	0.218	0.055	[0.112, 0.329]
T-PATE: Weighting-based Estimator	IPW	0.400	0.210	[-0.003, 0.79]
T-PATE: Weighting-based Estimator	wLS	0.254	0.090	[0.083, 0.437]
T-PATE: Outcome-based Estimator	OLS	0.256	0.161	[-0.049, 0.552]
T-PATE: Outcome-based Estimator	BART	0.138	0.096	[-0.041, 0.33]
T-PATE: Doubly Robust Estimator	AIPW with OLS	0.356	0.176	[0.006, 0.683]
T-PATE: Doubly Robust Estimator	AIPW with BART	0.353	0.173	[0.008, 0.692]
Measurement at +3 Weeks				
SATE	OLS	0.179	0.059	[0.059, 0.289]
T-PATE: Weighting-based Estimator	IPW	0.530	0.227	[0.077, 0.969]
T-PATE: Weighting-based Estimator	wLS	0.223	0.066	[0.095, 0.354]
T-PATE: Outcome-based Estimator	OLS	0.256	0.151	[-0.039, 0.544]
T-PATE: Outcome-based Estimator	BART	0.116	0.089	[-0.071, 0.287]
T-PATE: Doubly Robust Estimator	AIPW with OLS	0.320	0.159	[0.007, 0.64]
T-PATE: Doubly Robust Estimator	AIPW with BART	0.380	0.177	[0.033, 0.724]
Measurement at +6 Weeks				
SATE	OLS	0.263	0.056	[0.146, 0.37]
T-PATE: Weighting-based Estimator	IPW	0.606	0.235	[0.134, 1.052]
T-PATE: Weighting-based Estimator	wLS	0.384	0.074	[0.243, 0.527]
T-PATE: Outcome-based Estimator	OLS	0.282	0.151	[0, 0.564]
T-PATE: Outcome-based Estimator	BART	0.206	0.096	[0.025, 0.385]
T-PATE: Doubly Robust Estimator	AIPW with OLS	0.369	0.167	[0.041, 0.706]
T-PATE: Doubly Robust Estimator	AIPW with BART	0.446	0.162	[0.142, 0.748]
Measurement at +3 Months				
SATE	OLS	0.259	0.061	[0.133, 0.374]
T-PATE: Weighting-based Estimator	IPW	0.581	0.219	[0.136, 1.011]
T-PATE: Weighting-based Estimator	wLS	0.314	0.064	[0.189, 0.436]
T-PATE: Outcome-based Estimator	OLS	0.219	0.153	[-0.096, 0.523]
T-PATE: Outcome-based Estimator	BART	0.128	0.096	[-0.065, 0.32]
T-PATE: Doubly Robust Estimator	AIPW with OLS	0.285	0.156	[-0.024, 0.589]
T-PATE: Doubly Robust Estimator	AIPW with BART	0.321	0.181	[-0.04, 0.649]

K.2 Results for Bisgaard (2019) analysis in Figure 8

For the external validity analysis of Bisgaard (2019), we test each hypothesis by considering C- and Y-validity together using a sign-generalization test. Replicating the results from the original author, for each k outcome (Y_k) and context (c) pair, we estimate the effect of treatment (T_i) running the following regression separately within each context:

$$Y_{kc} \sim T_{ic}$$

where all regressions are estimated using robust standard errors. The resulting p-values, presented in Table A5 are used in the partial conjunction test.

Table A5: Numeric Values for Sign-Generalization Test for Bisgaard (2019).

Threshold	Outcome Variation	Context Variation	P-value
(H1) Incumbent Supporters			
1	Argument Rating	United States	0.000
2	Argument Rating	United States	0.000
3	Open-Ended	United States	0.000
4	Argument Rating	United States	0.000
5	Close-Ended	United States	0.000
6	Argument Rating	United States	0.000
7	Argument Rating	United States	0.000
8	Open-Ended	Denmark	0.003
9	Close-Ended	Denmark	0.152
10	Close-Ended	Denmark	0.248
11	Argument Rating	United States	0.251
12	Open-Ended	Denmark	0.251
(H2) Opposition Supporters			
1	Argument Rating	United States	0.000
2	Open-Ended	United States	0.000
3	Argument Rating	United States	0.000
4	Argument Rating	United States	0.000
5	Close-Ended	United States	0.000
6	Argument Rating	United States	0.000
7	Argument Rating	United States	0.000
8	Argument Rating	United States	0.000
9	Open-Ended	Denmark	0.000
10	Close-Ended	Denmark	0.000
11	Open-Ended	Denmark	0.000
12	Close-Ended	Denmark	0.221

K.3 Results for Broockman and Kalla (2016) analysis in Figure A1

In Figure A1 we conduct the T-PATE analysis separately by canvasser identity to evaluate T-validity. The estimators are the same as those used for Figure 7, described in Section K.1, conducted separately by whether the randomly assigned canvasser self identifies as transgender. The resulting point estimates and intervals are presented in Table A6.

Table A6: Numeric Values for T-PATE Estimates for Broockman and Kalla (2016) by Canvasser Identity in Figure A1.

Time Period	Estimator	Estimate	SE	95% CI
All				
+3 Days	SATE	0.218	0.055	[0.112, 0.329]
+3 Days	T-PATE: Weighting-based Estimator	0.254	0.090	[0.083, 0.437]
+3 Weeks	SATE	0.179	0.059	[0.059, 0.289]
+3 Weeks	T-PATE: Weighting-based Estimator	0.223	0.066	[0.095, 0.354]
+6 Weeks	SATE	0.263	0.056	[0.146, 0.37]
+6 Weeks	T-PATE: Weighting-based Estimator	0.384	0.074	[0.243, 0.527]
+3 Months	SATE	0.259	0.061	[0.133, 0.374]
+3 Months	T-PATE: Weighting-based Estimator	0.314	0.064	[0.189, 0.436]
Non-Transgender Canvasser				
+3 Days	SATE	0.140	0.070	[0.005, 0.283]
+3 Days	T-PATE: Weighting-based Estimator	0.148	0.096	[-0.039, 0.34]
+3 Weeks	SATE	0.140	0.073	[0.001, 0.292]
+3 Weeks	T-PATE: Weighting-based Estimator	0.206	0.089	[0.027, 0.373]
+6 Weeks	SATE	0.235	0.071	[0.09, 0.368]
+6 Weeks	T-PATE: Weighting-based Estimator	0.366	0.088	[0.203, 0.542]
+3 Months	SATE	0.235	0.079	[0.076, 0.385]
+3 Months	T-PATE: Weighting-based Estimator	0.387	0.088	[0.218, 0.556]
Transgender Canvasser				
+3 Days	SATE	0.370	0.107	[0.152, 0.581]
+3 Days	T-PATE: Weighting-based Estimator	0.401	0.143	[0.104, 0.667]
+3 Weeks	SATE	0.248	0.110	[0.051, 0.471]
+3 Weeks	T-PATE: Weighting-based Estimator	0.276	0.145	[0.009, 0.567]
+6 Weeks	SATE	0.303	0.097	[0.116, 0.501]
+6 Weeks	T-PATE: Weighting-based Estimator	0.345	0.107	[0.137, 0.543]
+3 Months	SATE	0.380	0.125	[0.142, 0.636]
+3 Months	T-PATE: Weighting-based Estimator	0.393	0.125	[0.139, 0.615]

K.4 Results for Bisgaard (2019) analysis in Figure A2

For the external validity analysis for Bisgaard (2019) in Figure A2, we further evaluate contextual variation due to the Denmark ruling coalition changing over time. The p-values are identical to those in Section K.2, with the Denmark ruling coalition included as a separate value in "Context". The original author's replication code can be found at https://doi.org/10.7910/DVN/FTFJTV.

Table A7: Numeric Values for Sign-Generalization Test for Bisgaard (2019) in Figure A2.

Threshold	Outcome Variation	Context Variation	P-value
(H1) Incumbent Supporters			
1	Argument Rating	United States	0.000
2	Argument Rating	United States	0.000
3	Open-Ended	United States	0.000
4	Argument Rating	United States	0.000
5	Close-Ended	United States	0.000
6	Argument Rating	United States	0.000
7	Argument Rating	United States	0.000
8	Open-Ended	Denmark (Center-left)	0.003
9	Close-Ended	Denmark (Center-right)	0.152
10	Close-Ended	Denmark (Center-right)	0.248
11	Argument Rating	United States	0.251
12	Open-Ended	Denmark (Center-right)	0.251
(H2) Opposition Supporters			
1	Argument Rating	United States	0.000
2	Open-Ended	United States	0.000
3	Argument Rating	United States	0.000
4	Argument Rating	United States	0.000
5	Close-Ended	United States	0.000
6	Argument Rating	United States	0.000
7	Argument Rating	United States	0.000
8	Argument Rating	United States	0.000
9	Open-Ended	Denmark (Center-left)	0.000
10	Close-Ended	Denmark (Center-right)	0.000
11	Open-Ended	Denmark (Center-right)	0.000
12	Close-Ended	Denmark (Center-right)	0.221

K.5 Results for Young (2019) analysis in Figure A3

For the external validity analysis for Young (2019), we test each hypothesis across outcome and treatment variations. The resulting p-values are combined using a sign-generalization test. Replicating the results from the original author, for each k outcome (Y_k) and treatment (t) pair, we estimate the effect of treatment (T_t) using the following weighted least squares regression:

$$Y_k \sim T_t$$

where the weights are the inverse probability weights from the block-assignment. Standard errors are estimated using robust standard errors. The resulting p-values, presented in Table A8, are used in the partial conjunction test. The original author's replication code can be found at https://doi.org/10.7910/DVN/UNNCTR.

Table A8: Numeric Values for Sign-Generalization Test for Young (2019) in Figure A3.

Hypothesis	Threshold	Treatment Variation	Outcome Variation	P-value
(H1)	1	Political Fear	Survey (Current)	0.000
(H1)	2	Political Fear	Survey (Current)	0.000
(H1)	3	Political Fear	Survey (Current)	0.000
(H1)	4	Political Fear	Survey (Current)	0.000
(H1)	5	Political Fear	Survey (Future)	0.000
(H1)	6	Political Fear	Survey (Future)	0.000
(H1)	7	Political Fear	Survey (Current)	0.000
(H1)	8	Political Fear	Survey (Current)	0.000
(H1)	9	Political Fear	Survey (Future)	0.000
(H1)	10	General Fear	Survey (Current)	0.000
(H1)	11	Political Fear	Survey (Future)	0.000
(H1)	12	General Fear	Survey (Current)	0.000
(H1)	13	Political Fear	Survey (Future)	0.000
(H1)	14	General Fear	Survey (Current)	0.000
(H1)	15	General Fear	Survey (Current)	0.000
(H1)	16	Political Fear	Survey (Future)	0.000
(H1)	17	General Fear	Survey (Current)	0.000
(H1)	18	General Fear	Survey (Future)	0.000
(H1)	19	General Fear	Survey (Current)	0.000
(H1)	20	General Fear	Survey (Future)	0.000
(H1)	21	General Fear	Survey (Future)	0.000
(H1)	22	General Fear	Survey (Future)	0.000
(H1)	23	General Fear	Survey (Future)	0.000

Table A8: Numeric Values for Sign-Generalization Test for Young (2019) in Figure A3. *(continued)*

Hypothesis	Threshold	Treatment Variation	Outcome Variation	P-value
(H1)	24	General Fear	Survey (Future)	0.000
(H1)	25	Political Fear	Behavioral	0.000
(H1)	26	General Fear	Behavioral	0.011
(H2)	1	Political Fear	Survey (Current)	0.000
(H2)	2	Political Fear	Survey (Current)	0.000
(H2)	3	Political Fear	Survey (Current)	0.000
(H2)	4	Political Fear	Survey (Future)	0.000
(H2)	5	Political Fear	Survey (Future)	0.000
(H2)	6	Political Fear	Survey (Current)	0.000
(H2)	7	Political Fear	Survey (Future)	0.000
(H2)	8	Political Fear	Survey (Future)	0.000
(H2)	9	Political Fear	Survey (Current)	0.000
(H2)	10	Political Fear	Survey (Current)	0.001
(H2)	11	Political Fear	Survey (Future)	0.004
(H2)	12	Political Fear	Survey (Future)	0.019
(H2)	13	General Fear	Survey (Current)	0.087
(H2)	14	General Fear	Survey (Current)	0.101
(H2)	15	General Fear	Survey (Current)	0.136
(H2)	16	General Fear	Survey (Future)	0.189
(H2)	17	General Fear	Survey (Current)	0.189
(H2)	18	General Fear	Survey (Current)	0.189
(H2)	19	General Fear	Survey (Current)	0.236
(H2)	20	General Fear	Survey (Future)	0.236
(H2)	21	General Fear	Survey (Future)	0.236
(H2)	22	General Fear	Survey (Future)	0.236
(H2)	23	General Fear	Survey (Future)	0.282
(H2)	24	General Fear	Survey (Future)	0.363
(H3)	1	Political Fear	Survey (Future)	0.000
(H3)	2	Political Fear	Survey (Future)	0.000
(H3)	3	General Fear	Survey (Current)	0.000
(H3)	4	Political Fear	Survey (Future)	0.001
(H3)	5	Political Fear	Survey (Current)	0.002
(H3)	6	Political Fear	Survey (Current)	0.004
(H3)	7	General Fear	Survey (Future)	0.005
(H3)	8	General Fear	Survey (Future)	0.006

Table A8: Numeric Values for Sign-Generalization Test for Young (2019) in Figure A3. *(continued)*

Hypothesis	Threshold	Treatment Variation	Outcome Variation	P-value
(H3)	9	General Fear	Survey (Future)	0.010
(H3)	10	Political Fear	Survey (Current)	0.010
(H3)	11	Political Fear	Survey (Current)	0.010
(H3)	12	General Fear	Survey (Current)	0.023
(H3)	13	Political Fear	Survey (Future)	0.023
(H3)	14	General Fear	Survey (Future)	0.028
(H3)	15	Political Fear	Survey (Future)	0.030
(H3)	16	Political Fear	Survey (Current)	0.033
(H3)	17	Political Fear	Survey (Current)	0.034
(H3)	18	Political Fear	Survey (Future)	0.034
(H3)	19	General Fear	Survey (Current)	0.118
(H3)	20	General Fear	Survey (Current)	0.119
(H3)	21	General Fear	Survey (Future)	0.119
(H3)	22	General Fear	Survey (Current)	0.174
(H3)	23	General Fear	Survey (Current)	0.174
(H3)	24	General Fear	Survey (Future)	0.174

K.6 Results for Dunning et al. (2019) analysis in Figure A4

We conduct a sign-generalization test of the results from Dunning et al. (2019) in Figure A4. To construct the p-values we use the point estimates and standard errors presented in the original paper. The resulting p-values presented in Table A9.

Table A9: Numeric Values for Sign-generalization test for Dunning et al. (2019) in Figure A4.

Threshold	Subgroup Variation	Context Variation	P-value
(H1) Vote Choice			
1	Good News	Uganda 1	0.897
2	Bad News	Mexico	1.000
3	Bad News	Uganda 2	1.000
4	Good News	Uganda 2	1.000
5	Good News	Brazil	1.000
6	Bad News	Brazil	1.000
7	Good News	Benin	1.000
8	Bad News	Benin	1.000
9	Good News	Burkina Faso	1.000
10	Bad News	Uganda 1	1.000
11	Good News	Mexico	1.000
12	Bad News	Burkina Faso	1.000
(H2) Turnout			
1	Good News	Uganda 1	0.176
2	Good News	Brazil	1.000
3	Good News	Uganda 2	1.000
4	Bad News	Uganda 1	1.000
5	Bad News	Uganda 2	1.000
6	Bad News	Brazil	1.000
7	Good News	Benin	1.000
8	Bad News	Benin	1.000
9	Good News	Burkina Faso	1.000
10	Good News	Mexico	1.000
11	Bad News	Burkina Faso	1.000
12	Bad News	Mexico	1.000

K.7 Results for Dehejia, Pop-Eleches and Samii (2021) analysis in Figure A5

We conduct a sign-generalization test of the results from Dehejia, Pop-Eleches and Samii (2021) in Figure A5. To construct the p-values we use the point estimates and standard errors presented in the original paper. The original analysis can be found at https://doi.org/10.6084/m9.figshare.8794991.v in Appendix Table 1. The resulting p-values presented in Table A10.

Table A10: Numeric Values for Sign-generalization test for Dehejia, Pop-Eleches and Samii (2021) in Figure A5.

Threshold	Decade	Year of Census	Income Group	Country	P-value
Outcome: Have More Kids					
1	1980	1980	Upper middle income	Argentina	0.000
2	1990	1991	Upper middle income	Argentina	0.000
3	2000	2001	Upper middle income	Argentina	0.000
4	2000	2001	Upper middle income	Armenia	0.000
5	1990	1991	Upper middle income	Brazil	0.000
6	1970	1970	Upper middle income	Brazil	0.000
7	2000	2000	Upper middle income	Brazil	0.000
8	1980	1980	Upper middle income	Brazil	0.000
9	1990	1998	Lower middle income	Cambodia	0.000
10	1990	1992	High income	Chile	0.000
11	1990	1990	Upper middle income	China	0.000
12	1980	1982	Upper middle income	China	0.000
13	1990	1993	Upper middle income	Colombia	0.000
14	1980	1985	Upper middle income	Colombia	0.000
15	2000	2005	Upper middle income	Colombia	0.000
16	2000	2002	Upper middle income	Cuba	0.000
17	1970	1975	High income	France	0.000
18	1960	1968	High income	France	0.000
19	1990	1999	High income	France	0.000
20	1960	1962	High income	France	0.000
21	1990	1990	High income	France	0.000
22	1980	1982	High income	France	0.000
23	1980	1981	High income	Greece	0.000
24	1990	1991	High income	Greece	0.000
25	2000	2001	High income	Greece	0.000
26	1970	1971	High income	Greece	0.000
27	1990	1990	High income	Hungary	0.000

Table A10: Numeric Values for Sign-generalization test for Dehejia, Pop-Eleches and Samii (2021) in Figure A5. *(continued)*

Threshold	Decade	Year of Census	Income Group	Country	P-value
28	1980	1980	High income	Hungary	0.000
29	1990	1999	Lower middle income	India	0.000
30	1990	1999	Lower middle income	Kyrgyz Republic	0.000
31	1990	1990	Upper middle income	Mexico	0.000
32	2000	2000	Upper middle income	Mexico	0.000
33	1990	1993	Upper middle income	Peru	0.000
34	2000	2007	Upper middle income	Peru	0.000
35	1990	1995	Lower middle income	Philippines	0.000
36	2000	2000	Lower middle income	Philippines	0.000
37	1990	1990	Lower middle income	Philippines	0.000
38	2000	2002	High income	Romania	0.000
39	1990	1992	High income	Romania	0.000
40	1970	1977	High income	Romania	0.000
41	1990	1996	Upper middle income	South Africa	0.000
42	1990	1991	High income	Spain	0.000
43	2000	2001	High income	Spain	0.000
44	2000	2000	Upper middle income	Thailand	0.000
45	1990	1990	Upper middle income	Thailand	0.000
46	1990	1991	High income	United Kingdom	0.000
47	2000	2005	High income	United States	0.000
48	2000	2000	High income	United States	0.000
49	1980	1980	High income	United States	0.000
50	1990	1990	High income	United States	0.000
51	1970	1970	High income	United States	0.000
52	1960	1960	High income	United States	0.000
53	1980	1989	Lower middle income	Vietnam	0.000
54	1990	1999	Lower middle income	Vietnam	0.000
55	2000	2001	High income	Austria	0.000
56	2000	2001	High income	Italy	0.000
57	1980	1981	High income	Austria	0.000
58	1990	1990	Upper middle income	Ecuador	0.000
59	2000	2001	Upper middle income	South Africa	0.000
60	1960	1960	Upper middle income	Brazil	0.000
61	1990	1991	High income	Austria	0.000
62	2000	2001	Lower middle income	Nepal	0.000

Table A10: Numeric Values for Sign-generalization test for Dehejia, Pop-Eleches and Samii (2021) in Figure A5. *(continued)*

Threshold	Decade	Year of Census	Income Group	Country	P-value
63	1990	1999	Upper middle income	Belarus	0.000
64	1980	1982	High income	Chile	0.000
65	2000	2001	Upper middle income	Ecuador	0.000
66	1980	1980	Upper middle income	Thailand	0.000
67	1970	1973	Upper middle income	Colombia	0.000
68	2000	2002	High income	Chile	0.000
69	1980	1981	High income	Portugal	0.000
70	1990	1993	Lower middle income	India	0.000
71	1970	1970	High income	Chile	0.000
72	1990	1998	Lower middle income	Pakistan	0.000
73	1970	1970	Upper middle income	Argentina	0.000
74	1980	1984	Upper middle income	Costa Rica	0.000
75	1980	1980	High income	Switzerland	0.000
76	2000	2000	Lower middle income	Mongolia	0.000
77	1970	1970	High income	Hungary	0.000
78	1990	1990	High income	Switzerland	0.000
79	2000	2000	Upper middle income	Costa Rica	0.000
80	1970	1971	High income	Austria	0.000
81	2000	2001	High income	Hungary	0.000
82	1980	1980	High income	Puerto Rico	0.000
83	1980	1987	Lower middle income	India	0.000
84	1990	1990	High income	Panama	0.000
85	1990	1995	Upper middle income	Mexico	0.000
86	2000	2000	Upper middle income	Malaysia	0.000
87	1970	1970	Upper middle income	Thailand	0.000
88	1990	1990	High income	Puerto Rico	0.000
89	1990	1997	Upper middle income	Iraq	0.000
90	2000	2000	High income	Puerto Rico	0.000
91	2000	2000	High income	Switzerland	0.001
92	1970	1972	High income	Israel	0.001
93	1980	1982	Upper middle income	Ecuador	0.001
94	2000	2000	High income	Panama	0.002
95	2000	2007	Upper middle income	South Africa	0.002
96	2000	2001	High income	Portugal	0.003
97	2000	2004	Upper middle income	Jordan	0.004

Table A10: Numeric Values for Sign-generalization test for Dehejia, Pop-Eleches and Samii (2021) in Figure A5. *(continued)*

Threshold	Decade	Year of Census	Income Group	Country	P-value
98	1990	1992	Lower middle income	Bolivia	0.006
99	1980	1980	Upper middle income	Malaysia	0.009
100	1990	1991	Upper middle income	Malaysia	0.011
101	1980	1983	High income	Israel	0.013
102	1980	1983	Lower middle income	India	0.017
103	2000	2001	Lower middle income	Bolivia	0.017
104	1990	1991	High income	Portugal	0.023
105	1970	1976	Lower middle income	Bolivia	0.205
106	2000	2005	High income	Puerto Rico	0.248
107	1970	1970	Upper middle income	Mexico	0.281
108	1970	1974	Upper middle income	Ecuador	0.371
109	1990	1999	Lower middle income	Kenya	0.409
110	1990	1998	Low income	Mali	0.534
111	1970	1970	Upper middle income	Malaysia	0.581
112	1970	1970	High income	Switzerland	0.643
113	1980	1980	High income	Panama	0.955
114	1980	1989	Lower middle income	Mongolia	1.000
115	1990	1995	High income	Israel	1.000
116	1990	1996	Low income	Guinea	1.000
117	1980	1983	Low income	Guinea	1.000
118	1970	1970	High income	Panama	1.000
119	1980	1988	Lower middle income	Senegal	1.000
120	1970	1973	Lower middle income	Pakistan	1.000
121	2000	2002	Low income	Rwanda	1.000
122	2000	2002	High income	Slovenia	1.000
123	1970	1970	High income	Puerto Rico	1.000
124	1960	1960	High income	Panama	1.000
125	2000	2002	Lower middle income	Tanzania	1.000
126	2000	2000	Lower middle income	Ghana	1.000
127	1980	1987	Low income	Mali	1.000
128	1990	1991	Low income	Rwanda	1.000
129	2000	2002	Lower middle income	Senegal	1.000
130	2000	2002	Low income	Uganda	1.000
131	1970	1973	Upper middle income	Costa Rica	1.000
132	1980	1989	Lower middle income	Kenya	1.000

Table A10: Numeric Values for Sign-generalization test for Dehejia, Pop-Eleches and Samii (2021) in Figure A5. *(continued)*

Threshold	Decade	Year of Census	Income Group	Country	P-value
133	1990	1991	Low income	Uganda	1.000
134	1980	1988	Lower middle income	Tanzania	1.000
Outcome: Economically Active					
1	1990	1990	High income	United States	0.000
2	1980	1980	High income	United States	0.000
3	1980	1982	Upper middle income	China	0.001
4	1980	1982	High income	France	0.002
5	2000	2000	High income	United States	0.013
6	1990	1990	High income	Hungary	0.019
7	2000	2001	Upper middle income	Armenia	0.154
8	1990	1991	High income	United Kingdom	0.203
9	1980	1987	Lower middle income	India	0.215
10	1990	1990	Upper middle income	China	0.425
11	1980	1981	High income	Austria	0.955
12	2000	2002	Low income	Rwanda	1.000
13	2000	2001	High income	Portugal	1.000
14	1990	1990	Lower middle income	Philippines	1.000
15	1990	1990	High income	France	1.000
16	1980	1980	High income	Switzerland	1.000
17	1990	1991	Upper middle income	Argentina	1.000
18	1990	1999	High income	France	1.000
19	1990	1991	High income	Austria	1.000
20	1990	1990	Upper middle income	Mexico	1.000
21	2000	2000	Upper middle income	Malaysia	1.000
22	2000	2001	Lower middle income	Nepal	1.000
23	1960	1960	High income	United States	1.000
24	1970	1971	High income	Greece	1.000
25	1970	1970	High income	United States	1.000
26	2000	2001	High income	Hungary	1.000
27	1990	1999	Lower middle income	Kyrgyz Republic	1.000
28	1960	1962	High income	France	1.000
29	1990	1999	Upper middle income	Belarus	1.000
30	1990	1999	Lower middle income	Vietnam	1.000
31	1960	1968	High income	France	1.000
32	2000	2000	High income	Puerto Rico	1.000

Table A10: Numeric Values for Sign-generalization test for Dehejia, Pop-Eleches and Samii (2021) in Figure A5. *(continued)*

Threshold	Decade	Year of Census	Income Group	Country	P-value
33	2000	2001	High income	Austria	1.000
34	1990	1991	Upper middle income	Malaysia	1.000
35	1980	1989	Lower middle income	Vietnam	1.000
36	2000	2000	Upper middle income	Brazil	1.000
37	1980	1983	Low income	Guinea	1.000
38	1980	1980	Upper middle income	Malaysia	1.000
39	2000	2002	High income	Chile	1.000
40	1990	1991	Upper middle income	Brazil	1.000
41	2000	2001	Upper middle income	Argentina	1.000
42	2000	2000	Lower middle income	Ghana	1.000
43	1970	1970	Upper middle income	Brazil	1.000
44	2000	2002	Lower middle income	Tanzania	1.000
45	1970	1973	Lower middle income	Pakistan	1.000
46	1990	1991	Low income	Rwanda	1.000
47	1990	1991	Low income	Uganda	1.000
48	1980	1980	Upper middle income	Argentina	1.000
49	1980	1989	Lower middle income	Kenya	1.000
50	1990	1990	High income	Switzerland	1.000
51	2000	2002	High income	Slovenia	1.000
52	1980	1983	Lower middle income	India	1.000
53	1990	1991	High income	Greece	1.000
54	2000	2000	Upper middle income	Mexico	1.000
55	1970	1970	High income	Switzerland	1.000
56	1990	1992	High income	Romania	1.000
57	1990	1993	Lower middle income	India	1.000
58	1970	1970	Upper middle income	Argentina	1.000
59	1970	1975	High income	France	1.000
60	2000	2002	Lower middle income	Senegal	1.000
61	1990	1999	Lower middle income	Kenya	1.000
62	2000	2001	High income	Spain	1.000
63	1990	1991	High income	Spain	1.000
64	1980	1980	High income	Panama	1.000
65	1980	1985	Upper middle income	Colombia	1.000
66	1970	1972	High income	Israel	1.000
67	2000	2002	Upper middle income	Cuba	1.000

Table A10: Numeric Values for Sign-generalization test for Dehejia, Pop-Eleches and Samii (2021) in Figure A5. *(continued)*

Threshold	Decade	Year of Census	Income Group	Country	P-value
68	2000	2001	Upper middle income	Ecuador	1.000
69	1980	1988	Lower middle income	Senegal	1.000
70	2000	2005	High income	Puerto Rico	1.000
71	1990	1993	Upper middle income	Peru	1.000
72	1980	1981	High income	Greece	1.000
73	1960	1960	Upper middle income	Brazil	1.000
74	2000	2000	High income	Switzerland	1.000
75	1990	1995	Upper middle income	Mexico	1.000
76	1990	1990	High income	Panama	1.000
77	2000	2005	High income	United States	1.000
78	1980	1980	Upper middle income	Brazil	1.000
79	1990	1997	Upper middle income	Iraq	1.000
80	1970	1970	High income	Panama	1.000
81	2000	2001	Lower middle income	Bolivia	1.000
82	1990	1996	Upper middle income	South Africa	1.000
83	1980	1984	Upper middle income	Costa Rica	1.000
84	1990	1999	Lower middle income	India	1.000
85	2000	2001	High income	Greece	1.000
86	1980	1982	High income	Chile	1.000
87	1990	1992	High income	Chile	1.000
88	2000	2000	Upper middle income	Costa Rica	1.000
89	1970	1971	High income	Austria	1.000
90	2000	2002	Low income	Uganda	1.000
91	1980	1981	High income	Portugal	1.000
92	1980	1982	Upper middle income	Ecuador	1.000
93	2000	2001	High income	Italy	1.000
94	2000	2007	Upper middle income	Peru	1.000
95	1970	1976	Lower middle income	Bolivia	1.000
96	1970	1970	Upper middle income	Malaysia	1.000
97	2000	2002	High income	Romania	1.000
98	1970	1970	Upper middle income	Mexico	1.000
99	2000	2007	Upper middle income	South Africa	1.000
100	2000	2000	Lower middle income	Mongolia	1.000
101	1990	1991	High income	Portugal	1.000
102	2000	2005	Upper middle income	Colombia	1.000

Table A10: Numeric Values for Sign-generalization test for Dehejia, Pop-Eleches and Samii (2021) in Figure A5. *(continued)*

Threshold	Decade	Year of Census	Income Group	Country	P-value
103	2000	2004	Upper middle income	Jordan	1.000
104	1980	1987	Low income	Mali	1.000
105	1970	1973	Upper middle income	Costa Rica	1.000
106	1980	1988	Lower middle income	Tanzania	1.000
107	1990	1998	Lower middle income	Cambodia	1.000
108	1990	1996	Low income	Guinea	1.000
109	2000	2001	Upper middle income	South Africa	1.000
110	1990	1990	High income	Puerto Rico	1.000
111	1990	1992	Lower middle income	Bolivia	1.000
112	1990	1993	Upper middle income	Colombia	1.000
113	1970	1973	Upper middle income	Colombia	1.000
114	1990	1990	Upper middle income	Ecuador	1.000
115	1990	1998	Low income	Mali	1.000
116	1960	1960	High income	Panama	1.000
117	2000	2000	High income	Panama	1.000
118	1970	1970	High income	Chile	1.000
119	1990	1995	High income	Israel	1.000
120	1970	1974	Upper middle income	Ecuador	1.000

K.8 Results for Bisbee et al. (2017) analysis in Figure A6

We conduct a sign-generalization test of the results from Bisbee et al. (2017) in Figure A6. To construct the p-values we use the point estimates and standard errors presented in the original paper. The original analysis can be found at https://www.journals.uchicago.edu/doi/epdf/10.1086/691280 in Table A1. The resulting p-values presented in Table A11.

Table A11: Numeric Values for Sign-generalization test for Bisbee et al. (2017) in Figure A6.

Threshold	Decade	Year of Census	Income Group	Country	P-value
Outcome: Economically Active					
1	1980	1980	High income	United States	0.000
2	1990	1990	High income	United States	0.000
3	1980	1982	High income	France	0.001
4	1980	1982	Upper middle income	China	0.001
5	2000	2000	High income	United States	0.005
6	1990	1990	High income	Hungary	0.025
7	1990	1990	Upper middle income	China	0.244
8	2000	2001	Upper middle income	Armenia	0.309
9	1990	1999	Lower middle income	Kyrgyz Republic	0.431
10	2000	2001	High income	Hungary	0.960
11	1990	1991	Upper middle income	Argentina	1.000
12	1980	1987	Lower middle income	India	1.000
13	1990	1990	High income	France	1.000
14	1990	1990	Lower middle income	Philippines	1.000
15	1960	1960	High income	United States	1.000
16	1990	1990	Upper middle income	Mexico	1.000
17	2000	2001	High income	Portugal	1.000
18	1970	1970	High income	United States	1.000
19	1970	1971	High income	Greece	1.000
20	1980	1980	High income	Switzerland	1.000
21	1990	1999	High income	France	1.000
22	2000	2000	Upper middle income	Malaysia	1.000
23	1960	1962	High income	France	1.000
24	1990	1999	Upper middle income	Belarus	1.000
25	2000	2001	Lower middle income	Nepal	1.000
26	1960	1968	High income	France	1.000
27	1990	1999	Lower middle income	Vietnam	1.000
28	1990	1991	Upper middle income	Brazil	1.000
29	2000	2000	High income	Puerto Rico	1.000

Table A11: Numeric Values for Sign-generalization test for Bisbee et al. (2017) in Figure A6. *(continued)*

Threshold	Decade	Year of Census	Income Group	Country	P-value
30	2000	2001	Upper middle income	Argentina	1.000
31	1980	1989	Lower middle income	Vietnam	1.000
32	1990	1991	Upper middle income	Malaysia	1.000
33	2000	2002	High income	Chile	1.000
34	2000	2000	Upper middle income	Brazil	1.000
35	1980	1980	Upper middle income	Malaysia	1.000
36	1970	1970	Upper middle income	Brazil	1.000
37	1980	1980	Upper middle income	Argentina	1.000
38	1990	1991	High income	Greece	1.000
39	2000	2002	Low income	Rwanda	1.000
40	1990	1993	Lower middle income	India	1.000
41	2000	2000	Upper middle income	Mexico	1.000
42	1990	1990	High income	Switzerland	1.000
43	1970	1970	High income	Switzerland	1.000
44	2000	2002	Lower middle income	Tanzania	1.000
45	1980	1988	Lower middle income	Tanzania	1.000
46	1980	1983	Lower middle income	India	1.000
47	1970	1975	High income	France	1.000
48	1980	1983	Low income	Guinea	1.000
49	2000	2002	High income	Slovenia	1.000
50	1980	1980	High income	Panama	1.000
51	1970	1970	Upper middle income	Argentina	1.000
52	1990	1992	High income	Romania	1.000
53	2000	2001	High income	Spain	1.000
54	1970	1972	High income	Israel	1.000
55	2000	2001	Upper middle income	Ecuador	1.000
56	2000	2002	Upper middle income	Costa Rica	1.000
57	2000	2005	High income	Puerto Rico	1.000
58	1990	1991	High income	Spain	1.000
59	1980	1985	Upper middle income	Colombia	1.000
60	1980	1980	Upper middle income	Brazil	1.000
61	1990	1995	Upper middle income	Mexico	1.000
62	1990	1997	Upper middle income	Iraq	1.000
63	1980	1988	Lower middle income	Senegal	1.000
64	1960	1960	Upper middle income	Brazil	1.000

Table A11: Numeric Values for Sign-generalization test for Bisbee et al. (2017) in Figure A6. *(continued)*

Threshold	Decade	Year of Census	Income Group	Country	P-value
65	2000	2005	High income	United States	1.000
66	1990	1993	Upper middle income	Peru	1.000
67	1980	1987	Low income	Mali	1.000
68	1980	1981	High income	Greece	1.000
69	2000	2001	Lower middle income	Bolivia	1.000
70	1990	1996	Upper middle income	South Africa	1.000
71	1990	1999	Lower middle income	India	1.000
72	2000	2000	High income	Switzerland	1.000
73	1990	1990	High income	Panama	1.000
74	1980	1984	Upper middle income	Costa Rica	1.000
75	2000	2002	Low income	Uganda	1.000
76	1970	1970	High income	Panama	1.000
77	1990	1992	High income	Chile	1.000
78	1980	1982	High income	Chile	1.000
79	2000	2002	Lower middle income	Senegal	1.000
80	2000	2001	High income	Greece	1.000
81	2000	2000	Upper middle income	Costa Rica	1.000
82	2000	2007	Upper middle income	Peru	1.000
83	1970	1970	Upper middle income	Malaysia	1.000
84	1980	1982	Upper middle income	Ecuador	1.000
85	1980	1981	High income	Portugal	1.000
86	2000	2001	High income	Italy	1.000
87	1970	1976	Lower middle income	Bolivia	1.000
88	1970	1973	Upper middle income	Costa Rica	1.000
89	1990	1991	High income	Portugal	1.000
90	2000	2002	High income	Romania	1.000
91	2000	2007	Upper middle income	South Africa	1.000
92	1970	1970	Upper middle income	Mexico	1.000
93	2000	2000	Lower middle income	Mongolia	1.000
94	2000	2005	Upper middle income	Colombia	1.000
95	2000	2004	Upper middle income	Jordan	1.000
96	1980	1989	Lower middle income	Kenya	1.000
97	1990	1998	Lower middle income	Cambodia	1.000
98	1990	1995	High income	Israel	1.000

Table A11: Numeric Values for Sign-generalization test for Bisbee et al. (2017) in Figure A6. *(continued)*

Threshold	Decade	Year of Census	Income Group	Country	P-value
99	1990	1996	Low income	Guinea	1.000
100	1990	1991	Low income	Uganda	1.000
101	1960	1960	High income	Panama	1.000
102	1990	1990	High income	Puerto Rico	1.000
103	1990	1993	Upper middle income	Colombia	1.000
104	2000	2001	Upper middle income	South Africa	1.000
105	1990	1992	Lower middle income	Bolivia	1.000
106	1990	1998	Low income	Mali	1.000
107	2000	2000	Lower middle income	Ghana	1.000
108	1970	1973	Upper middle income	Colombia	1.000
109	2000	2000	High income	Panama	1.000
110	1990	1990	Upper middle income	Ecuador	1.000
111	1970	1974	Upper middle income	Ecuador	1.000
112	1970	1970	High income	Chile	1.000

References

Benjamini, Yoav and Ruth Heller. 2008. "Screening for Partial Conjunction Hypotheses." *Biometrics* 64(4):1215–1222.

Bisbee, James, Rajeev Dehejia, Cristian Pop-Eleches and Cyrus Samii. 2017. "Local Instruments, Global Extrapolation: External Validity of the Labor Supply–Fertility Local Average Treatment Effect." *Journal of Labor Economics* 35(S1):S99–S147.

Bisgaard, Martin. 2019. "How Getting the Facts Right Can Fuel Partisan-Motivated Reasoning." *American Journal of Political Science* 63(4):824–839.

Blair, Graeme, Jasper Cooper, Alexander Coppock, Macartan Humphreys and Luke Sonnet. 2019. *estimatr: Fast Estimators for Design-Based Inference*. R package version 0.20.0. **URL:** *https://CRAN.R-project.org/package=estimatr*

Broockman, David and Joshua Kalla. 2016. "Durably Reducing Transphobia: A Field Experiment On Door-to-Door Canvassing." *Science* 352(6282):220–224.

Dehejia, Rajeev, Cristian Pop-Eleches and Cyrus Samii. 2021. "From Local to Global: External Validity in a Fertility Natural Experiment." *Journal of Business & Economic Statistics* 39(1):217–243.

Deville, Jean-Claude and Carl-Erik Särndal. 1992. "Calibration Estimators in Survey Sampling." *Journal of the American Statistical Association* 87(418):376–382.

Dunning, Thad, Guy Grossman, Macartan Humphreys, Susan D Hyde et al. 2019. "Voter Information Campaigns and Political Accountability: Cumulative Findings from a Preregistered Meta-Analysis of Coordinated Trials." *Science Advances* 5(7):eaaw2612.

Hainmueller, Jens. 2012. "Entropy Balancing For Causal Effects: A Multivariate Reweighting Method to Produce Balanced Samples in Observational Studies." *Political Analysis* 20(1):25–46.

Hartman, Erin, Richard Grieve, Roland Ramsahai and Jasjeet S Sekhon. 2015. "From Sample Average Treatment Effect to Population Average Treatment Effect on the Treated." *Journal of the Royal Statistical Society. Series A (Statistics in Society)* 178(3):757–778.

Hill, Jennifer L. 2011*a*. "Bayesian Nonparametric Modeling for Causal Inference." *Journal of Computational and Graphical Statistics* 20(1):217–240.

Hill, Jennifer L. 2011*b*. "Bayesian Nonparametric Modeling for Causal Inference." *Journal of Computational and Graphical Statistics* 20(1):217–240.

Lin, Winston. 2013. "Agnostic Notes on Regression Adjustments to Experimental Data: Reexamining Freedman's Critique." *The Annals of Applied Statistics* 7(1):295–318.

Lumley, Thomas. 2020. "survey: analysis of complex survey samples.". R package version 4.0.

Tsiatis, Anastasios. 2006. *Semiparametric Theory and Missing Data*. Springer.

Young, Lauren E. 2019. "The Psychology of State Repression: Fear and Dissent Decisions in Zimbabwe." *American Political Science Review* 113(1):140–155.