Online Supplementary Appendix

Elements of External Validity:
Framework, Design, and Analysis

Table of Contents

In the Online Supplementary Appendix II, we provide additional details as follows.

A Effect-Generalization

We examine identification and 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.

A.1 Identification of the T-PATE

Assumption A1 (Identification Assumptions for X- and C-validity)

- Contextual Exclusion Restriction: For all $t \in \mathcal{T}$, $\mathbf{m} \in \mathcal{M}$, and all units,

$$\begin{aligned} & Y_i(T=1, \mathbf{M}=\mathbf{m}, c) - Y_i(T=0, \mathbf{M}=\mathbf{m}, c) \\ = \ & Y_i(T=1, \mathbf{M}=\mathbf{m}, c^*) - Y_i(T=0, \mathbf{M}=\mathbf{m}, c^*), \end{aligned} \qquad (1)$$

  where $\mathbf{M}$ are context-moderators as defined in Section 3.2.4. $\mathcal{T}$ is the support of the treatment variable T and $\mathcal{M}$ is the support of the context moderators $\mathbf{M}$.

- Ignorability of Sampling and Treatment Effect Heterogeneity: For all $\mathbf{x} \in \mathcal{X}$, $\mathbf{m} \in \mathcal{M}$,

$$Y_i(T=1, \mathbf{M}=\mathbf{m}) - Y_i(T=0, \mathbf{M}=\mathbf{m}) \perp\!\!\!\perp S_i \mid \mathbf{X}_i = \mathbf{x}, \mathbf{M}_i = \mathbf{m}, C_i = c \qquad (2)$$

$$Y_i(T=1, \mathbf{M}=\mathbf{m}) - Y_i(T=0, \mathbf{M}=\mathbf{m}) \perp\!\!\!\perp C_i \mid \mathbf{X}_i = \mathbf{x}, \mathbf{M}_i = \mathbf{m}, \qquad (3)$$

  where $\mathcal{X}$ is the support of the pre-treatment covariates $\mathbf{X}$.

- Overlap: For all $\mathbf{x} \in \mathcal{X}$, $\mathbf{m} \in \mathcal{M}$,

$$0 < \Pr(S_i = 1 \mid \mathbf{X}_i = \mathbf{x}, \mathbf{M}_i = \mathbf{m}, C_i = c) < 1 \qquad (4)$$

$$0 < \Pr(C_i = c \mid \mathbf{X}_i = \mathbf{x}, \mathbf{M}_i = \mathbf{m}) < 1 \qquad (5)$$

$$0 < \Pr(C_i = c^* \mid \mathbf{X}_i = \mathbf{x}, \mathbf{M}_i = \mathbf{m}) < 1 \qquad (6)$$

- Consistency: For all units,

$$Y_i \ = \ Y_i(T = T_i, \mathbf{M} = \mathbf{M}_i) \qquad (7)$$

Theorem A1 (Identification of the T-PATE under X- and C-validity)
Under Assumption A1 and the randomization of treatment assignment in experiments, the T-PATE is identified as follows.

$$\begin{aligned} & \mathbb{E}_{\mathcal{P}^*}[Y_i(T=1, c^*) - Y_i(T=0, c^*)] \\ = \ & \sum_{\mathbf{m} \in \mathcal{M}} \sum_{\mathbf{x} \in \mathcal{X}} \{\mathbb{E}(Y_i \mid T_i = 1, S_i = 1, C_i = c, \mathbf{M}_i = \mathbf{m}, \mathbf{X}_i = \mathbf{x}) \\ & \quad - \mathbb{E}(Y_i \mid T_i = 0, S_i = 1, C_i = c, \mathbf{M}_i = \mathbf{m}, \mathbf{X}_i = \mathbf{x})\} \Pr(\mathbf{M}_i = \mathbf{m}, \mathbf{X}_i = \mathbf{x} \mid C_i = c^*), \end{aligned}$$

where the sum may be interpreted as integral when appropriate.

Proof. In this proof, for notational simplicity, we use $Y_i(1, \mathbf{m})$ and $Y_i(0, \mathbf{m})$ instead of $Y_i(T = 1, \mathbf{M} = \mathbf{m})$ and $Y_i(T = 0, \mathbf{M} = \mathbf{m})$.

$$
\begin{aligned}
& \mathbb{E}_{\mathcal{P}^*}[Y_i(T=1,c^*) - Y_i(T=0,c^*)] \\
= \; & \sum_{\mathbf{m}\in\mathcal{M}} \sum_{\mathbf{x}\in\mathcal{X}} \mathbb{E}\{Y_i(T=1,c^*) - Y_i(T=0,c^*) \mid \mathbf{M}_i = \mathbf{m}, \mathbf{X}_i = \mathbf{x}, C_i = c^*\} \Pr(\mathbf{M}_i = \mathbf{m}, \mathbf{X}_i = \mathbf{x} \mid C_i = c^*) \\
= \; & \sum_{\mathbf{m}\in\mathcal{M}} \sum_{\mathbf{x}\in\mathcal{X}} \mathbb{E}\{Y_i(1,\mathbf{m}) - Y_i(0,\mathbf{m}) \mid \mathbf{M}_i = \mathbf{m}, \mathbf{X}_i = \mathbf{x}, C_i = c^*\} \Pr(\mathbf{M}_i = \mathbf{m}, \mathbf{X}_i = \mathbf{x} \mid C_i = c^*) \\
= \; & \sum_{\mathbf{m}\in\mathcal{M}} \sum_{\mathbf{x}\in\mathcal{X}} \mathbb{E}\{Y_i(1,\mathbf{m}) - Y_i(0,\mathbf{m}) \mid C_i = c, \mathbf{M}_i = \mathbf{m}, \mathbf{X}_i = \mathbf{x}\} \Pr(\mathbf{M}_i = \mathbf{m}, \mathbf{X}_i = \mathbf{x} \mid C_i = c^*) \\
= \; & \sum_{\mathbf{m}\in\mathcal{M}} \sum_{\mathbf{x}\in\mathcal{X}} \mathbb{E}\{Y_i(1,\mathbf{m}) - Y_i(0,\mathbf{m}) \mid S_i = 1, C_i = c, \mathbf{M}_i = \mathbf{m}, \mathbf{X}_i = \mathbf{x}\} \Pr(\mathbf{M}_i = \mathbf{m}, \mathbf{X}_i = \mathbf{x} \mid C_i = c^*) \\
= \; & \sum_{\mathbf{m}\in\mathcal{M}} \sum_{\mathbf{x}\in\mathcal{X}} \big[\mathbb{E}\{Y_i(1,\mathbf{m}) \mid T_i = 1, S_i = 1, C_i = c, \mathbf{M}_i = \mathbf{m}, \mathbf{X}_i = \mathbf{x}\} \\
& \qquad - \mathbb{E}\{Y_i(0,\mathbf{m}) \mid T_i = 0, S_i = 1, C_i = c, \mathbf{M}_i = \mathbf{m}, \mathbf{X}_i = \mathbf{x}\}\big] \Pr(\mathbf{M}_i = \mathbf{m}, \mathbf{X}_i = \mathbf{x} \mid C_i = c^*), \\
= \; & \sum_{\mathbf{m}\in\mathcal{M}} \sum_{\mathbf{x}\in\mathcal{X}} \{\mathbb{E}(Y_i \mid T_i = 1, S_i = 1, C_i = c, \mathbf{M}_i = \mathbf{m}, \mathbf{X}_i = \mathbf{x}) \\
& \qquad - \mathbb{E}(Y_i \mid T_i = 0, S_i = 1, C_i = c, \mathbf{M}_i = \mathbf{m}, \mathbf{X}_i = \mathbf{x})\} \Pr(\mathbf{M}_i = \mathbf{m}, \mathbf{X}_i = \mathbf{x} \mid C_i = c^*),
\end{aligned}
$$

where the first equality follows from the definition of the T-PATE and the rules of conditional probability, the second from the contextual exclusion restriction (equation (1) in Assumption A1), the third from the conditional ignorability of the selection into contexts (equation (3) in Assumption A1), and the fourth from the conditional ignorability of the selection into experiments (equation (2) in Assumption A1). The fifth inequality follows from the randomization of treatment assignment within the experiment, which implies

$$
\{Y_i(1,\mathbf{m}), Y_i(0,\mathbf{m})\} \perp\!\!\!\perp T_i \mid S_i = 1, C_i = c, \mathbf{M}_i = \mathbf{m}, \mathbf{X}_i = \mathbf{x}, \tag{8}
$$

for all $\mathbf{x} \in \mathcal{X}$, $\mathbf{m} \in \mathcal{M}$. Note that, as we emphasize in Section 3.2.4, it is critical that both context-moderators $\mathbf{M}_i$ and covariates used for the X-validity $\mathbf{X}_i$ are pre-treatment, that is, not affected the treatment variable (Rosenbaum, 1984). The final sixth equality follows from the consistency of the potential outcomes (equation (7) in Assumption A1). This completes the proof. □

A.2 Three Classes of Estimators

Here we provide the formal expressions of the three classes of the T-PATE estimators. We prove their statistical properties in Appendix H in the Online Supplementary Appendix II. $\widehat{\pi}_i$ and $\widehat{\theta}_i$ are defined in Section 5.2.

A.2.1 Weighting-based Estimator

Inverse Probability Weighted (IPW) estimator:

$$
\widehat{\tau}_{\texttt{IPW}} \equiv \frac{\sum_{i=1}^R \widehat{\theta}_i \widehat{\pi}_i \delta_i \mathbf{1}\{C_i = c\} S_i T_i Y_i}{\sum_{i=1}^R \widehat{\theta}_i \widehat{\pi}_i \delta_i \mathbf{1}\{C_i = c\} S_i T_i} - \frac{\sum_{i=1}^R \widehat{\theta}_i \widehat{\pi}_i (1-\delta_i) \mathbf{1}\{C_i = c\} S_i (1-T_i) Y_i}{\sum_{i=1}^R \widehat{\theta}_i \widehat{\pi}_i (1-\delta_i) \mathbf{1}\{C_i = c\} S_i (1-T_i)}, \tag{9}
$$

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).

Weighted Least Squares:

$$(\widehat{\alpha}, \widehat{\tau}_{\texttt{wLS}}, \widehat{\gamma}) = \underset{\alpha,\tau,\gamma}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} w_i (Y_i - \alpha - \tau T_i - \mathbf{Z}_i^{\top} \gamma)^2 \tag{10}$$

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

A.2.2 Outcome-based 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

$$\widehat{g}_t(\mathbf{X}_j, \mathbf{M}_j) \quad \equiv \quad \widehat{\mathbb{E}}(Y_i \mid T_i = t, \mathbf{M}_j, \mathbf{X}_j, S_i = 1, C_i = c).$$

A.2.3 Doubly Robust Estimator

Augmented Inverse Probability Weighted (AIPW) estimator:

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

where we use R to denote the sum of the sample size in the experiment (n) and in the target population data (N).

A.3 Inference with Bootstrap

To compute standard errors, we rely on the nonparametric bootstrap (Efron and Tibshirani, 1994). In particular, we consider the bootstrap over experimental samples. If randomization is done with block or cluster randomization, we also incorporate such treatment assignment mechanisms. While the target population data is often considered fixed, it is also possible to bootstrap over the target population data to account for population sampling uncertainty.

B Sign-Generalization

B.1 Fisher's Combined p-value

In some applications, researchers can obtain p-values that are independent across variations. For example, when researchers run experiments across multiple contexts, experimental data across context are independent and thus, p-values are independent. In such cases, researchers can use the Fisher's method to combine p-values and compute the partial conjunction p-value (Benjamini and Heller, 2008). For the partial conjunction null hypothesis $\widetilde{H}_0^r$, the partial-conjunction p-value is

$$\widetilde{p}_{(r)} = \Pr\left(\chi^2_{2(K-r+1)} \geq -2\sum_{i=r}^{K} \log p_{(i)}\right).$$

B.2 Statistical Power and Purposive Variations

One key consideration is the number of purposive variations to include. On the one hand, the larger number of purposive variations increases the credibility of sign-generalization because the required range assumption is more tenable. On the other hand, a larger number of purposive variations usually leads to smaller effective sample sizes and larger standard errors. In particular, for T- and C-validity, introducing more variations means smaller sample size for each treatment level and each context.

In general, researchers should prioritize the credibility of sign-generalization and incorporate enough purposive variations to satisfy the range assumption. This is because sign-generalization becomes impossible without sufficient purposive variations, whereas there are several ways to mitigate concerns about standard errors. In particular, researchers can supplement the design of purposive variations with methods that improve statistical efficiency, such as blocking and the design-based method of using pre-treatment variables (see e.g., Gerber and Green, 2012), as usually recommended in any experimental analyses.

C Empirical Applications: Full Analysis

We apply the proposed methodologies to the three empirical applications described in Section 2. In this section of the supplementary material, we provide additional discussion and analyses for the three studies.

C.1 Field Experiment: Reducing Transphobia

In Section 7.1, we discussed effect-generalization for Broockman and Kalla (2016). In this section, we provide additional implementation details for the described estimators. We also discuss T- and C-validity within the context of this experiment.

C.1.1 Effect-Generalization: Estimation Details

To estimate the T-PATE, we adjust for age, sex, race/ethnicity, ideology, religiosity, and partisan identification, which include all variables measured in both the experiment and the CCES.[1] We focus on the estimation of the intent-to-treat effect in the target population, defined using the CCES data of respondents from Florida (Ansolabehere and Schaffner, 2017) with

[1] In the experiment, the authors used age, sex, and race/ethnicity as reported on the voter file. There may be some measurement differences compared to the self-reported measures used in the CCES. The remainder of the variables used the same question, although we collapsed responses to common values across the two datasets. Age is measured using a five-category age bucket for weighting, and age in years for BART. Race/ethnicity is

validation from Enamorado and Imai (2018). We estimate the T-PATE using three classes of estimators we discussed in Section 5.1. Weighting-based estimators include IPW and weighted least squares with the control variables pre-specified in the original authors' pre-analysis plan. Sampling weights are estimated via calibration (Deville and Särndal, 1992; Hartman et al., 2015), which matches weighted marginals of the experimental sample to the target population marginals with a max weight of 10. For the outcome-based estimators, we use OLS and a more flexible model, BART (Hill, 2011). Finally, we implement two doubly robust estimators; the AIPW with OLS and the AIPW with BART, as described in Section 5.1, where the weights are estimated using calibration. We use function `tpate` in our forthcoming `R` package `evalid` to implement all estimators.

C.1.2 Y-validity

In addition to the measurement of outcomes over time, Broockman and Kalla (2016)'s study improves Y-validity in a number of ways. First, they measure outcomes in surveys ostensibly unrelated to the intervention. While not easily quantifiable, this helps increase external validity of the measure by avoiding survey satisficing among respondents aware of the intervention. Second, typical of modern field experiments, Broockman and Kalla (2016) measure a variety of survey questions on attitudes toward transgender people, which jointly approximate real-world attitudes. We follow the original analysis that combines multiple outcomes into a single index. In particular, we estimate the impact on this index 3 days, 3 weeks, 6 weeks, and 3 months after the intervention. These multiple outcome variations can also be used to conduct the sign-generalization test described in Section 6 under much weaker assumptions. An example of this approach is discussed in our reanalysis of Bisgaard (2019) and Young (2019).

C.1.3 T-validity

The intervention used in Broockman and Kalla (2016) is a complex, compound treatment. The authors note "we cannot be certain that perspective-taking is responsible for any effects or that active processing is responsible for their duration; being primarily concerned with external validity and seeking to limit suspicion, we did not probe intervening processes or restrict the scope of the conversations as a laboratory study would" (p. 222). This implies the target treatment is the whole canvassing interaction, not merely the perspective-taking aspect.

Individuals were randomly assigned to receive a door-to-door canvassing intervention from either a self-identified transgender or non-transgender individual, who revealed their identity during the intervention. This provides an opportunity to evaluate one aspect of T-validity. Having a conversation with a self-identified transgender individual may have a different effect

coded as a three-level category for "Black," "Hispanic," and "White/Other." Ideology and partisanship are coded as seven-point scales, and religiosity is a five-level factor. Indicators are created for factors in regression and weighting methods, and are entered as ordered categories for the causal BART. In the supplementary material of Broockman and Kalla (2016), the original authors compared a subset of the above six variables, {age, sex, and race/ethnicity}, of the experimental sample with those of all voters in Miami-Dade county using the voter file.

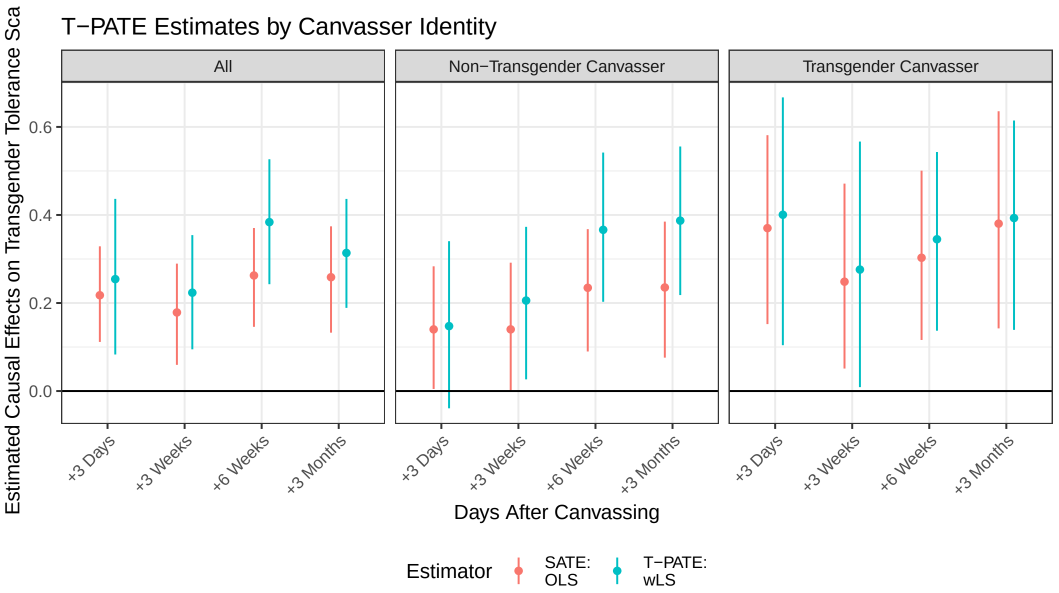

Figure A1: T-PATE Estimates for Broockman and Kalla (2016) By Canvasser Identity *Note:* The x-axis within panels represents survey waves (3 days, 3 weeks, 6 weeks, 3 months). Panels present canvasser identity. Estimates are for the SATE, with pre-specified controls (pink) and the T-PATE with weighted least squares (blue).

than a conversation with a non-transgender individual. The authors partnered with an LGBT organization, where about a quarter of the canvassers self-identified as transgender, a much larger proportion than the general population, and one that may be infeasible in a larger-scale intervention. Therefore, researchers may be interested in whether the treatment is robust to partnerships with organizations with a different distribution of canvasser identity in which fewer individuals identify themselves as transgender.

Figure A1 presents the T-PATE estimates by canvasser identity and time-period. The SATE estimate (pink) and the T-PATE estimate based on the weighted least squares estimator (blue), both with pre-specified controls, are positive across canvasser identity and time-period, and the T-PATE estimates are similar to the SATE estimates. This suggests that the intervention can have similar effects even after considering three dimensions together, i.e., X-, Y-, and T-validity. It is important to re-emphasize that no formal analysis can guarantee "full" external validity, and we should be clear about the targets of external validity. This analysis provides evidence for (1) X-validity with all adults in Florida under Assumption 1, (2) Y-validity over three months, and (3) T-validity with respect to the identity of canvassers.

C.1.4 C-validity

As is common in field experiments, the authors conducted their analysis in one geography, Miami. Therefore, it is difficult to evaluate C-validity in terms of geography. However, the authors discuss one important aspect of context that could impact the effectiveness of the intervention, noting that "[a]ttack ads featuring antitransgender stereotypes are another common feature of political campaigns waged in advance of public votes on nondiscrimination laws" (p. 223). This contextual variable, the ad environment, might change how the treatment affects outcomes, which is the C-validity question. To address this concern, they evaluate support for the Miami-Dade anti-discrimination law during each post-treatment survey wave. During wave three, to "examine whether support for the law would withstand such [negative attack ads], we showed subjects one of three such ads from recent political campaigns elsewhere, then immediately asked about the law again" (p. 223). They note that, while support for the law decreases in response to the attack ad, individuals subjected to the perspective-taking intervention were still more positive towards the law than those in the control group. The negative impact of the ad on support for the law diminished by wave 4.

We use a sign-generalization test to evaluate C-validity of the results across the pre- and post-attack ad measurement in wave 3, as well as the measurement in wave 4.[2] The target context here is one in which negative attack ads are present during the canvassing period. The pre-ad measurement likely has a larger effect than might be present in a context with a large negative ad campaign, whereas the post-ad measurement, taken directly after viewing an attack ad, likely represents a stronger impact of a negative ad campaign, giving credence to the range assumption. The measurement in wave 4 is likely somewhere in the middle, given the time since the individual viewed the attack ad.

We first focus on C-validity together with Y-validity. To do so, we consider an OLS estimator with pre-specified controls without sampling weights (i.e., we are not considering X-validity for now). We find that the point estimates of the intervention effect are all positive, and using the partial conjunction test, we find that all outcomes across three-time periods have a p-value that is significant at the $\alpha = 0.05$ level.

We then evaluate three dimensions together, C-, Y-, and X-validity. In this analysis, since the focus is on a law in Miami-Dade county, we address X-validity by weighting to a target population defined by the full list of registered voters from which the experimental sample was drawn.[3] We incorporated estimated sampling weights to a weighted least squares estimator we described in Section 5. Using the partial conjunction test, while the point estimates are all positive and consistent with the theory, no estimate rejects the conventional significance level at any threshold. Therefore, there is limited evidence that the intervention has the same positive effects across different ad environments among all Miami-Dade voters.[4]

[2] The authors note in their original analysis that the term "transgender" had not been defined for the control group in the first and second waves, mitigating the effect of the intervention. Therefore, we focus on the later waves where "transgender" is defined for all subjects.

[3] Weighting is done using all available voter file characteristics, including sex, race/ethnicity, age, turnout in 2010, 2012, and 2014, and party registration.

[4] We note that, in the original manuscript, the authors focus on the complier average causal effect, which was statistically significant at the $\alpha = 0.05$ level in a one-tailed test for each of the measurements described.

C.1.5 Cost-Benefit Analysis

Effect-generalization is most useful for randomized experiments that have policy implications because cost-benefit considerations will be affected by the actual effect size. While a formal cost-benefit analysis is beyond the scope of this paper, we discuss a simple approach to cost-benefit analysis and clarify how the T-PATE estimate will affect such analyses.[5]

We use b_i to represent unit i's benefit corresponding to a one unit change in the outcome of interest, and use c_i to represent the cost of the treatment for unit i. These parameters b_i and c_i depend on the application, and thus, we keep them general here. This generality is important because different organizations will have different costs and gain different benefits from the same intervention. The average utility of the intervention to the target population can be written as $\frac{1}{N}\sum_{i=1}^{N}(\tau_i b_i - c_i)$ where τ_i is the treatment effect for unit i. When the average utility is positive, researchers may argue that the intervention is cost-effective.

Suppose the benefit parameter b_i is constant across units, denoted by b. Then, the average utility can be simplified to be $b \times \textsf{T-PATE} - \frac{1}{N}\sum_{i=1}^{N} c_i$. Therefore, the T-PATE estimate is directly useful for the cost-benefit analysis. In particular, we can estimate the average utility by $b \times \widehat{\textsf{T-PATE}} - \frac{1}{N}\sum_{i=1}^{N} c_i$ where $\widehat{\textsf{T-PATE}}$ is estimated by one of the three classes of estimators discussed in Section 5. The standard error is $b \times \widehat{\textsf{se}}(\widehat{\textsf{T-PATE}})$ where $\widehat{\textsf{se}}(\widehat{\textsf{T-PATE}})$ is the standard error of the T-PATE estimator estimated using the bootstrap. Therefore, researchers can test whether the average utility is statistically significantly different from zero.

Therefore, even though our analysis of the T-PATE showed point estimates similar to the SATE, our analysis revealed that standard errors of the T-PATE estimate are larger than those of the SATE estimate. This suggests that statistical uncertainty for the average utility of the treatment are often larger when researchers appropriately take into account external validity concerns and conduct effect-generalization.

Finally, when benefit parameter b_i differs across subgroups, researchers can estimate the T-PATE separately for subgroups and apply the same logic. While formal cost benefit analysis has been rare in political science, future work can incorporate it into the potential outcomes framework and connect it more thoroughly to the question of external validity.

C.2 Survey Experiment: Partisan-Motivated Reasoning

In Section 7.2, we discussed a sign-generalization test for Bisgaard (2019) focusing on Y- and C-validity. We discuss X- and T-validity in this section.

C.2.1 X-validity

The studies, conducted by YouGov, are population-based surveys of the voting-age population. Population-based survey experiments are intended to be representative of the target population, increasing the likelihood of X-validity. The analyses in the original manuscript do not incorporate survey weights[6]; however, as noted in footnote 1 of the original manuscript,

[5] We thank an anonymous reviewer for encouraging us to examine the cost benefit analysis more.

[6] Weights are not available in the replication file.

YouGov used an "Active Sampling" technique for Studies 1-3, in which respondents are invited continuously to match "key characteristics of the target population" (p. 828). We conduct unweighted analyses here, and our target population is the same as the sample Bisgaard (2019) focused on.

C.2.2 T-validity

In each study, individuals are randomly assigned to read about a positive or negative change in GDP, or assigned to a control group in studies 1 and 2. To the degree possible, the only difference in the prompts is whether the change in GDP is cast in a positive or negative light. The target treatment is the provision of positive or negative economic information in everyday life, such as when reading news articles. The treatment is designed to "[keep] in touch with reality" while also "relatively strong and unambiguous to create a situation in which both stripes of partisans would acknowledge the facts at hand" (p. 828), indicating that the treatment effect considered within this experiment might be stronger than what we would observe in the real world. In this experiment, unfortunately, there is only one treatment implemented, and therefore, there is no purposive variation we can use for sign-generalization. If we can incorporate several treatments with varying degrees of reality, we can use the proposed sign-generalization test to evaluate this aspect of the T-validity.

C.2.3 C-validity

We considered the main contextual variations across the United States and Denmark in Section 7.2. Here, we consider an additional contextual variation available in the study. Another source of contextual variation occurs within Denmark, where the ruling party changes from a center-left to a center-right coalition between Studies 2 and 3. Therefore, if Bisgaard (2019)'s theory holds, those who support a center-left coalition would attribute responsibility to the government in the face of positive economic information in Study 2 (as supporters of the incumbent party), but the same people would attribute little responsibility to the government in the face of positive economic information in Study 3 (now as supporters of the opposition party).

Figure A2 presents the analysis from Section 7.2 of the main text, including the additional contextual variation of the Denmark ruling coalition. As can be seen, results for opposition supporters are strongest, including across the coalition variation in Denmark. However, the results from the Denmark center-right coalition do not support the hypothesis for incumbent supporters.

C.2.4 Discussion

The results in Section 7.2 suggest several important policy implications. First, as suggested in the original study, political campaigns emphasizing news about changes in GDP will likely have larger and more stable effects in the United States than in Denmark because the incumbent party's political responsibility for the economy is less clear, and the level of polarization among citizens is lower in Denmark than in the United States. Second, our new result based on

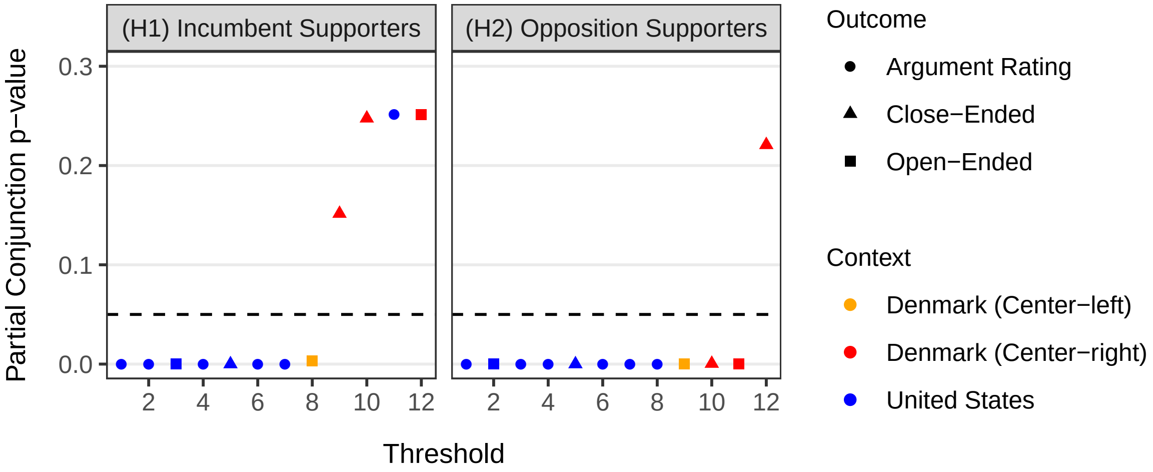

Figure A2: Sign-Generalization Test for Bisgaard (2019). *Note:* We combine causal estimates on multiple outcomes across four survey experiments in three contexts. Following Section 6, we report partial conjunction p-values for all thresholds.

the sign-generalization test suggests that such political campaigns work more effectively for opposition supporters, i.e., negative campaigns about the incumbent work better than positive campaigns.

C.3 Lab Experiment: The Effect of Emotions on Dissent in Autocracy

Young (2019) finds that fear plays a key role in shaping individuals' risk assessment of repression in an autocracy, which in turn affects the likelihood of dissent. We now consider how to conduct a formal external validity analysis of this theory. Like Bisgaard (2019), Young (2019)'s research question and hypotheses, which focus primarily on the direction of causal effects, fit well with sign-generalization. In particular, she formulates her main hypotheses as follows. Individuals in a state of fear will: (H1) express less dissent, (H2) be more pessimistic about the risk of repression, and (H3) be more pessimistic in their expectations of whether others will also dissent.[7]

We use the sign-generalization test to take into account T- and Y-validity together. We show how to conduct the sign-generalization test by combining variations in treatments and outcomes. We also discuss X- and C-validity.

C.3.1 T-validity

Young (2019) implemented two versions of the treatment in which participants were either directed to describe general fears, and directed away from experiences related to politics and elections (general fear condition), or they were directed to describe fears related to politics and elections (political fear condition). These two conditions are designed based on considerations of both preciseness and realism of treatments (see also Section 3.2.2). The general fear condi-

[7] Note that we focus our external validity analysis on the three main hypotheses listed above because no explicit purposive variation is available for the final fourth hypothesis (see p.142 of the original article).

tion is designed to be a “cleaner test of the effect of fear because in this condition participants are not even reflecting on information about repression that they already have,” and the political fear condition “more closely approximates the way that fear may be induced in practice in repressive environments, through memories or stories of brutal violence” (p. 144). These two treatment conditions are compared against a control condition, in which participants were asked to describe activities that make them feel relaxed.

Because the two treatments address both preciseness and realism, many interesting target treatments will satisfy the required range assumption (Assumption 5) under the purposive variations in Young (2019). We formally test whether causal estimates are consistently negative across variations in these treatment conditions. If we find the sign of causal estimates is stable, we can expect that a broad range of treatments inducing fear will also negatively affect the expressions of dissent.

C.3.2 Y-validity

For each of the hypotheses, Young (2019) measures a host of outcomes that are contextually relevant and span a range of risk levels. For the first hypothesis (H1), she measures six hypothetical acts (wearing an opposition party t-shirt, sharing a funny joke about the president, going to an opposition rally, refusing to go to a rally for the ruling party, telling a state security agent that she supports the opposition, and testifying in court against a perpetrator of violence) as well as one behavioral outcome (selecting a plastic wristband with a pro-democracy slogan vs. a non-political message). Similarly, for the second hypothesis (H2), measurements are taken to assess the likelihood individuals would experience six types of repression (threats, assault, destruction of property, sexual abuse, abduction, and murder) if they attended an opposition rally or meeting. Finally, for the third hypothesis (H3), she asks about the proportion of other opposition supporters that would engage in the six hypothetical acts of dissent from the first hypothesis. For each hypothetical attitude question, the respondents were also asked to evaluate the item for both the current period, when risks are lower, as well as around the next election, when risks are likely heightened.

The key is that these various questions were selected to cover a range of risky dissent behaviors. If the target outcome is a low-risk dissent behavior, it might be reasonable to assume that the purposive outcome variations in Young (2019) satisfy the required assumption (Assumption 5). However, some high-risk dissent behaviors are unlikely to range with the purposive variations. We take a conservative approach, and we interpret the sign-generalization test only with respect to low-risk dissent behaviors.

C.3.3 Sign-Generalization Test

For each hypothesis, we combine purposive variations for T-validity and Y-validity (see Table A1 for a summary). We have 2 (treatments) × 13 (outcomes) estimates for (H1), 2×12 for (H2), and 2×12 for (H3). We recode all outcomes such that each hypothesis predicts negative effects. We estimate effects using weighted least squares, accounting for the differential probability of treatment defined in the original analysis, and use HC2 robust standard errors as implemented

Hypothesis	Variations for T-Validity	Variations for Y-Validity
H1	General Fear, Political Fear, Control	Hypothetical acts of dissent (12) + Behavioral measure (1)
H2	General Fear, Political Fear, Control	Probability of experiencing different forms of repression (12)
H3	General Fear, Political Fear, Control	Proportion of other opposition supporters who will engage in hypothetical acts of dissent (12)

Table A1: Design of Purposive Variations for Young (2019).

in the `estimatr` package . Then, using the partial conjunction test (Section 6.2.2), we formally quantify the proportion of negative causal effects for each hypothesis. Given the number of comparisons is large for each hypothesis, the importance of employing the proposed approach and properly accounting for multiple comparisons is high.

Figure A3 presents the results from the partial conjunction tests for each hypothesis. We present the partial conjunction p-values for each threshold. Each p-value is colored by their treatment condition, with the general fear condition (green) and political fear condition (purple). The outcomes are represented by symbols, with the behavioral outcome presented as dots, survey questions assessed for the current period as triangles, and survey questions assessed for the future election as squares. For the first hypothesis, we find that 26 out of 26 outcomes (100%) have partial conjunction p-values less than the conventional significance level 0.05. There is strong evidence for the sign-generalizability of the first hypothesis (H1), that fear will reduce expressions of political dissent.

The evidence for the second and third hypotheses is more mixed. Young (2019) hypothesizes that people in a state of fear will be more pessimistic about the risk of repression in the second hypothesis (H2). We find that only 12 out of 24 outcomes (50%) have partial conjunction p-values less than 0.05, with support from the political fear condition but not from the general fear condition, indicating that a weaker treatment might not generalize. Regarding their belief about whether others will also engage in dissent (the third hypothesis), we find that the partial conjunction p-values are less than 0.05 for 18/24 (75%) of the outcomes, where again the political fear condition shows stronger support for the theory than the general fear condition. Therefore, there exists stronger evidence for the political fear treatment than for the general fear treatment.

C.3.4 C-validity

Young (2019)'s analysis does not provide a clear opportunity to test for context validity. The author notes that Zimbabwe has "a long history of repressive violence designed to reduce the political participation of opposition supporters" but that "when the study was carried out, active violence against opposition supporters was very low," which allowed for a context where individuals are in a repressive regime but did not require "exposing participants to unjustifiable risks" (p. 143). While the author does take hypothetical measures that prime different political contexts, asking if they would engage in dissent in the current time period as well as during the upcoming election, the measurements are not taken in different contexts.

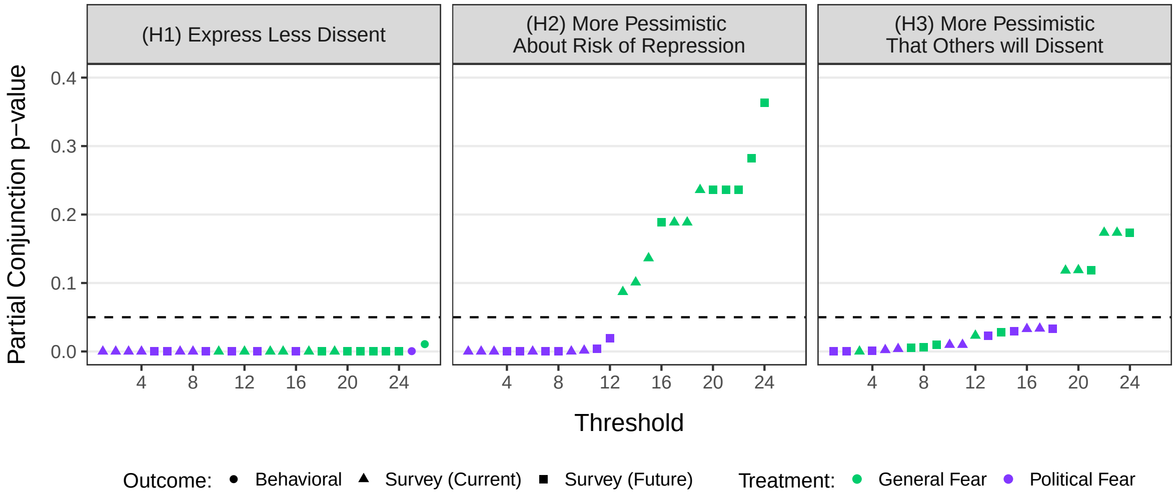

Figure A3: Sign-Generalization Test for Young (2019). *Note:* We combine causal estimates on multiple outcomes across two treatment-variations. In total, we have 26 causal estimates for the first hypothesis and 24 for the second and third hypotheses. Following Section 6, we report partial conjunction p-values for all thresholds.

To formally evaluate C-validity, future experiments can include purposive variations in context. For example, we can run one experiment close to an election and another far away from an election. This will induce variations in authoritarian pressure, which we can use to test the sign generalization in terms of C-validity. A multi-site experiment is a popular strategy to induce variations in geography, and we can assess whether causal effects are generalizable to other authoritarian regimes. The variations should be carefully chosen to meet the required range assumption for sign-generalization.

C.3.5 X-validity

In the appendix of the original paper, the author compares her sample to two nationally representative surveys across a number of important measures, including potential moderators. Overall, she finds her sample is representative across a number of measures, including gender, education, and many measures of victimization of pro-opposition individuals. She does find differences among poverty rates, as well as the number of pro-opposition individuals who reported that a family member had been killed for political reasons since 1980.

To account for some of the measurable differences, she conducts an analysis using an IPW estimator with post-stratification weights, to match the Afrobarometer on gender, age, education, and subjective measures of poverty. The resulting point estimates are very similar to the original analysis, indicating that concerns of X-validity are not impacting the results, under the assumption that the variables controlled for with the weights make sample selection and treatment effect heterogeneity conditionally independent (Assumption 1). She also conducted a sensitivity analysis on the number of strong opposition supporters in the sample, which cannot be accounted for in the weighting analysis. It also indicates that the results are robust to changes on this dimension, providing additional strength to the credibility.

D Metaketa

D.1 Motivating Example

Information about politician performance, such as their effectiveness and responsiveness, is an essential tool in democracy that can help voters hold politicians accountable and reduce corruption. Metaketa I (Dunning et al., 2019) aimed to study whether voter information campaigns, funded extensively by NGOs and nonprofits, are effective. The research team conducted a coordinated study with a common definition of treatment, in which the researchers worked with local partners to distribute "objective, nonpartisan performance information privately to individual voters within 2 months prior to the election" (p. 2) across five countries with harmonized baseline and outcome measures. This allows for cumulative learning and a replication of the same treatment across contexts; both valuable forms of external validity analysis. Ultimately, Dunning et al. (2019) find null effects of voter information campaigns on two outcomes of interest: vote choice, specifically voting for the incumbent, and voter turnout.

Strengths and Weaknesses for External Validity In many ways, Metaketas are designed to explicitly address the four dimensions of external validity. For example, in the Dunning et al. (2019) study, the inclusion of multiple, diverse sites improves both X- and C-validity. However, the sites are themselves not random draws of the units and contexts of theoretical interest, with a high concentration in the Global South. Effect-generalization to a new context, such as a country with strong ethnic divisions, still requires strong assumptions.

The common arm treatment bundles the types of information provision groups use in practice increases T-validity with respect to how information is commonly provided. However, pragmatic designs can limit T-validity for specific target-treatments of theoretical interest, such as public provision of information, or a information about politician actions vs outcomes.

One significant strength of the Metaketa is the harmonization of pre-treatment and outcome measures. The common measures across sites increase Y-validity, ensuring differences observed across sites are not attributable to different measurement strategies. However, coordination does not inherently ensure Y-validity if the measurements do not align with the target outcomes of interest, and the concerns we've outlined are still applicable. For example, we must still assume ignorable outcome variations if the target outcome is strength of support, or enthusiasm, for the incumbent, rather than the dichotomous vote for incumbent measured in the study.

D.2 Sign-Generalization Test

In their original analysis, Dunning et al. (2019) use a meta-analysis of their multi-site experiment to evaluate treatment effectiveness. The diversity of purposive variations on units, treatments, outcomes, and contexts measured in the Metaketa bolster the range assumption required for the sign-generalization test. We separately consider sign-generalization for the primary (H1) and secondary (H2) hypotheses listed in Section 3 of the supplementary materials for the original paper, which are reproduced below.[8]

(H1) Positive (negative) information increases (decreases) voter support for politicians.

(H2) Positive (negative) information increases (decreases) voter turnout.

Data For our re-analysis, we download the point estimates and standard errors from the meta-analysis model, retrieved from the authors' replication website (https://egap.shinyapps.io/metaketa_shiny/). This shiny app provides point estimates for all primary and intermediate outcomes.[9] We collect the estimates for the primary outcomes "vote for incumbent" and "voter turnout". The original study reports the effect of the treatment among two subgroups — those for whom the information provided exceeds prior beliefs on candidate performance (positive or "good news") or falls short of their baseline beliefs (negative or "bad news"), which we collect separately.

Analysis In our sign-generalization test, we consider two types of purposive variations, including country, addressing C-validity, and the "good" vs. "bad news" subgroup analysis, addressing X-validity. The range assumption requires we assume the target effect, for example for a country not included in the study such as Nigeria, lies within the effects seen in the five countries included in the study, and where the effect of the country-specific implementation of information provision lies within the good and bad news groups observed. We conduct the sign-generalization test separately for each hypothesis. This yields twelve estimates (6 sites × 2 subgroups) which we combine with a partial conjunction test for each outcome.

Results Figure A4 presents the results for the sign-generalization test for the theory presented in Dunning et al. (2019) that information provision affects vote choice (H1) and voter turnout (H2). The results indicate limited support for sign-generalizability; we cannot reject the null that none of the variations support the theory for either hypothesis. This is unsurprising in the context of the meta-analysis findings from the original study, which found only one statistically significant point estimate among the 24 estimates across contexts, subgroups, and outcomes. Note that we design the sign-generalization test to assess whether the treatment effect is positive or negative, as hypothesized in the original pre-analysis plan (H1 and H2 above), and we found that there is no evidence for either positive or negative causal effects. In this Metaketa I, an alternative interpretation of the experimental result is that the null effect is generalizable across six sites, which we could test with an appropriate equivalence test (Hartman and Hidalgo, 2018), while this should be considered as a post-hoc interpretation as the pre-analysis plan did not specify hypotheses in this way.

[8] We combine their component hypotheses (H1a and H1b; H2a and H2b) into a single hypothesis, respectively.

[9] We collect point estimates and clustered standard errors for each country using the following settings: we do not include covariate controls; we exclude non-contested elections in the Uganda 2 study (default); we include both LCV chairs and councilors in the Uganda 2 study (default); we weight each study equally (default).

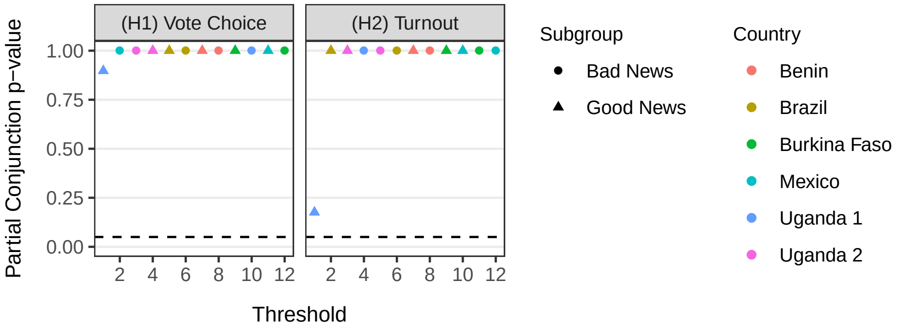

Figure A4: Sign-generalization test for Dunning et al. (2019) for the primary (vote for incumbent) and secondary (voter turnout) outcomes. Country is represented by color and subgroup by symbol.

D.3 Effect-Generalization

An alternative to sign-generalization would be to ask if the effect of voter information provision generalizes to a specific country outside of the six sites studied in the original trial, which we will refer to as the target country. As outlined in Section 3.2.4, many times when we consider generalizing to a different country we deal with both a change in the distribution of unit characteristics, leading to concerns about X-validity, as well as contextual moderators, leading to concerns of C-validity. We outline the steps a researcher can take to conduct such an analysis, following Figure 2 in the main text.

Dunning et al. (2019) took care to design a treatment that mimics common practice for information provision and relied on outcomes that are possible to measure in many target-countries. Therefore, we assume that concerns of T- and Y-validity are addressed by the design of the study, and consider the implemented treatment and outcome measures as our target-treatment and target-outcome measures, and focus on effect-generalization for X- and C-validity.

Step 1: Ask *whether* effect-generalization is possible. Recall that we must first evaluate whether the assumptions required for effect-generalization are justified. X-validity requires Assumption 1, which states that conditional on pre-treatment covariates, study participation and the individual level treatment effect are conditionally independent. In Dunning et al. (2019), the researchers collect a number of individual level characteristics in the baseline survey.[10] In order to conduct effect-generalization, a researcher should conduct a survey measuring these same variables, using the same measurement strategy, in the target country.

C-validity requires Assumption 4 which states that the causal effect for a given unit will be the same regardless of whether they are in the original study or in the target country, after

[10] This includes gender, age, coethic and cogender with the incumbent, years of education, relative wealth, incumbent party partisan attachment, vote history for last election, support for incumbent in last election, and baseline belief in incumbent party clientelism.

adjusting for context-moderators. To be plausible, we need to measure and adjust for context-moderators that capture how the causal effect differs across the experimental countries and the target country. In addition to possible individual level moderators, described above, Dunning et al. (2019), measure a number of contextual measures that might affect treatment effect heterogeneity.[11] The researcher should collect these, using the same measurement strategy, in their target country.

Step 2: Effect-Generalization (Estimate the T-PATE). After the researcher has carefully evaluated if these individual and contextual measures are likely to justify Assumptions 1 and 4, they can proceed to estimation of the T-PATE. This can be done with one of the three class of estimators described in Section 5.1, including weighting-based, outcome-based and doubly robust estimators (see extension to X- and C-validity together in Section 5.2). Which estimator is best depends on whether the researcher can accurately model the sampling or treatment effect heterogeneity processe (see Section 5.1.4). We generally suggest researchers use doubly robust estimators, which are consistent if either process is correctly specified, and the researcher need not know which one.

E External Validity Analysis of Observational Studies

E.1 Motivating Example

The role of fertility in women's labor-force participation is an important question for understanding the economic impacts of childbearing on family, and in particular, women's long term labor-force participation and success. However, isolating the effect of fertility is complicated by endogenous factors such as baseline female labor-force participation and fertility rate or culturally influenced delays in marriage and childbearing. Angrist and Evans (1998) use a natural experiment in which they note that families often have a preference for one child of each sex, allowing them to evaluate the impact of having two children of the same sex (referred to as the same-sex treatment) on third-child fertility decisions and labor-force participation of married women, aged 21-35 with children under 18, using U.S. census data from 1980 to 1990. They find significant negative effects of fertility on labor-force participation.

Natural Experiment We re-analyze two related studies that conduct an effect-generalization analysis of the impact of fertility on women's labor-force participation. We first consider Dehejia, Pop-Eleches and Samii (2021), who extend the original Angrist and Evans (1998) study to evaluate concerns about external validity using a world-wide dataset spanning the 1960 to 2010. This study relies on a natural experimental design in which the same-sex treatment is considered "as-if" randomly assigned. In their original analysis, the authors find that macro-level variables, including the proportion of educated mothers and the GDP of the country, are important for explaining treatment effect heterogeneity.

[11] This includes electoral competitiveness; whether the country uses a secret ballot; to what extent voters believe the country has free and fair elections; the Freedom House measure of freedom of the press; and the polity measure of democratic strength.

Instrumental Variables We also consider a study by Bisbee et al. (2017), who rely on the same dataset, but use the same-sex treatment as an instrument for fertility decisions (specifically, the decision to have a third child), and evaluate the impact on the labor-force participation; ultimately they find similar patterns of generalizability as Dehejia, Pop-Eleches and Samii (2021). These original studies each present an effect-generalization type analysis, therefore we focus on a sign-generalization analysis to complement the original findings.

E.2 Sign-Generalization

Data In our re-evaluation of Dehejia, Pop-Eleches and Samii (2021), we consider the sign-generalizability of the findings for the fertility outcome ("Have More Kids") as well as labor-force participation ("Economically Active"). We consider our analysis separately for each outcome measure and design. To evaluate the natural experiment, we collect the 254 point estimates and standard error estimates provided in Table A.1 of the original manuscript for each country-year-outcome dyad. Similarly, for the instrumental variables study we collect the 112 point estimates and standard error estimates from Table A.1 of Bisbee et al. (2017) for the "Economically Active" outcome evaluated in that study. We then use these estimates to calculate the one-sided p-value, which we input into our proposed sign-generalization test.

Analysis We consider purposive variations across two contextual variables. Dehejia, Pop-Eleches and Samii (2021) note that many countries, but not all, exhibit strong sex selectivity, especially for male children and that patterns have changed over time. They also evaluate the impact of macro-level variables, including gross domestic product. Based on these findings, our evaluation of sign-generalization, we consider purposive variations across geography and GDP, using the current World Bank income-group classification, and time, using the decade of the census.

When considering Assumption 5, the range assumption that justifies the sign-generalization test, we must assume that the effect in the target contexts lie within the convex hull of the observed purposive variations. While the original analysis covers 49 countries, it does not include estimates world wide, therefore when we ask if the results generalize to a specific country or year not included, we must assume the true effect is within the range of observed effects. There are still limitations to our study for other dimensions of external validity. For example, the authors limit their analysis to married women, aged 21-35 who have children under 18 at the time of the census. Therefore, to have X-validity, we must assume the effects are within the same convex hull for unwed or single, or older mothers, for whom the impact of fertility on labor-force participation may differ due to differing financial and familial support structures. For T-validity we either must focus on the same-sex treatment, or assume that a target treatment, such as the impact of a third child given two children of opposite sex, lies within the convex hull of the effects we have observed. These assumptions may be unreasonably strong given we observe no purposive variations for X and T.

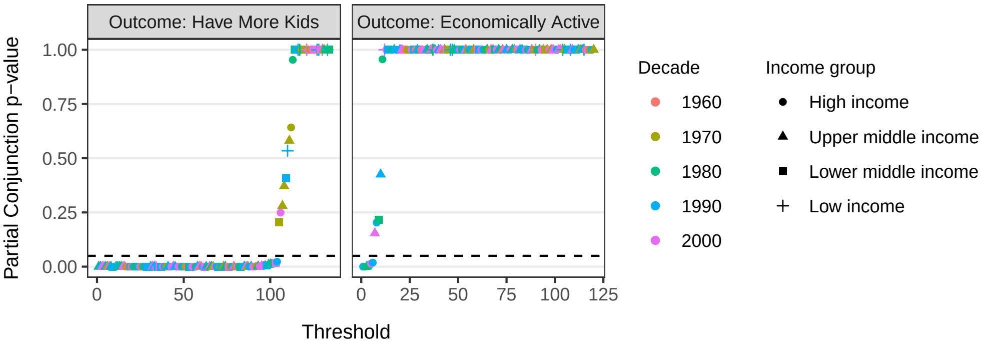

Figure A5: Sign-generalization test for Dehejia, Pop-Eleches and Samii (2021). Outcomes by study-design are represented by columns, country classification from the World Bank is represented by symbol, and color represents the decade of census.

E.2.1 Results for Natural Experiment Design

Figure A5 presents results of our sign-generalization test for the natural experiment conducted by Dehejia, Pop-Eleches and Samii (2021). Each panel represents a partial conjunction test conducted within the outcome of interest, with purposive variations across decade (differentiated by color) and income group (differentiated by symbol). We see that the results for the effect of our same-sex treatment on fertility ("Have more kids") demonstrate the strongest support for external validity. We can reject the null in favor of the alternative that at least 104 of our 134 estimates (78%) support the theory. However, the sign-generalization test indicates very little support for generalizability of the results the labor-force participation ("Economically Active"); this is unsurprising given most of the results were individually statistically insignificant in the original analysis.

Consistent with the original authors' finding that there is heterogeneity by country GDP, we find that the strongest evidence supporting the theory among high and upper middle income countries. In lower middle and low in come countries, the evidence is more mixed or does not provide statistical evidence supporting the theory.

When combined with the original authors' analysis, we see the value of both effect-generalization and sign-generalization. The thorough effect-generalization done in Dehejia, Pop-Eleches and Samii (2021) determines macro-level context moderators and micro-level sources of effect heterogeneity. Our sign-generalization analysis complements this by weakening the required identifying assumptions.

E.2.2 Results for Instrumental Variables Design

Figure A6 presents results of our sign-generalization test for instrumental variables design conducted by Bisbee et al. (2017). We represent purposive variations across decade (differentiated by color) and income group (differentiated by symbol). As with the reduced form analysis, the sign-generalization test indicates very little support for generalizability of the results for labor-force participation ("Economically Active"). We can only reject the null that at least 6

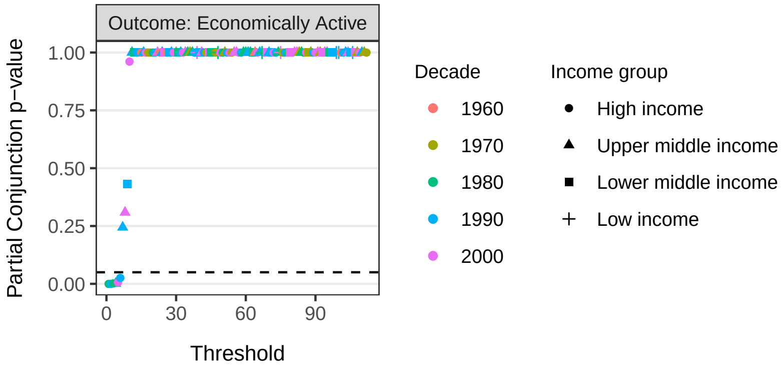

Figure A6: Sign-generalization test for Bisbee et al. (2017) for the "Economically Active" outcome. Country classification from the World Bank is represented by symbol, and color represents the decade of census.

of the results, out of 112, support the theory at the $\alpha = 0.05$ level; this is unsurprising given most of the results were individually statistically insignificant in the original analysis.

F Economics-Type Lab Experiment

F.1 General Discussion

In the main body of the paper, we focus on a lab-in-the-field example, Young (2019), which is rooted in the psychological style of lab experiments. Political scientists also rely on economics-style lab experiments. These experiments differ from psychology-style lab experiments on many important dimensions, including the incentives, design, and outcome measures (Bol, 2019; Dickson, 2011). Economics-style lab experiments tend to measures concrete outcome behaviors, including both individual and group behaviors, whereas psychology-style experiments tend to focus on individual reported attitudes. Economics-style experiments often rely on monetary incentives based on behavior instead of fixed compensation. The most important difference might be in the design of the experiment. Economics-style experiments tend to be more stylized and abstract, which gives the researcher more control over treatment and avoids confounding factors that exist outside of the lab, whereas psychology-style experiments emphasize realistic, and often bundled, treatments. In the following example, we consider the four dimensions of external validity for one economics-style experiment, Kanthak and Woon (2015).

F.2 Motivating Example

Legislatures in the United States from the local to the federal level exhibit a significant under-representation of female office holders. Kanthak and Woon (2015) contribute to a robust literature on factors that affect the decision of female legislators to run for office, and the barriers they face in becoming officeholders, by isolating the impact of election aversion in dissuading women from seeking office. Using a lab experiment conducted among undergraduate participants, researchers randomly assign a representative to be chosen among a pool of anonymous volunteers, or elected by plurality vote, to conduct an objective problem-solving task. They vary the private cost of running for election, as well as the electoral environment, which can be either truthful or strategic and prone to misinformation. Ultimately, the authors find that women are election averse — the fact that a representative is chosen by an election dissuades women from putting forth their name for consideration, holding all else equal — unless elections are both cost-less and completely truthful.

F.2.1 X-validity

A common criticism of lab experiments is their reliance on undergraduate participants. The authors argue "undergraduates are at similar life stages, not yet having embarked on their careers or started their families, and their youth and education should also make them less susceptible to gender-based social constraints on running for office" (p. 597). In order to address X-validity, for example when generalizing to a real-world electorate, we need to account for such factors by measuring and adjusting for pre-treatment covariates that make treatment effect heterogeneity conditionally independent of the sample selection process, or we must assume that these factors do not affect treatment effect heterogeneity.

F.2.2 T-validity

A common feature of economics-style lab experiments is their reliance on an abstract and stylized treatment. In Kanthak and Woon (2015), the laboratory setting allows the researchers to exert significant control over the experimental manipulation and therefore rule out common explanations for a woman's decision to run for office such as ability, risk preferences, and societal beliefs. The anonymous voting also limits the impact of women's perceptions about biases voters may hold. This allows the researchers to attribute the gender gap to the electoral context for deciding the representative (i.e. volunteer vs. election-based) and the associated costs.

While abstract treatments commonly used in economics-style lab experiments allow a researcher to isolate a single dimension of a complex treatment, it can affect T-validity. If our target-treatment is a real-world election, this treatment is bundled with the societal beliefs about women and personal risk preferences and ability, dimensions which might dwarf or exacerbate election aversion. To address T-validity, we must assume that the target real-world election treatment has the same effect as the effect of the anonymous electoral context treatment in the experiment.

F.2.3 Y-validity

Similar to field experiments, economics-style lab experiments often focus on behavioral outcomes, such as the decision to run for election, as opposed to elicited attitudes and preferences commonly used in survey and psychology-style lab experiments. While the focus on behavioral measures may be closer to target-outcomes, such as a decision to run for office in a real-world election, the local nature of the measurement in a hypothetical election game still requires that we must assume the difference between the experimental outcome and the target outcome is ignorable.

F.2.4 C-validity

The abstract, collaborative interactions of many economics-style lab experiments may impact context validity. For example, Kanthak and Woon (2015) rely on anonymous, computer-based interactions to limit the biases women may experience when deciding whether to run for office, and they focus on an objective problem-solving task with no gendered difference in demonstrated success. However, women who decide to run in real-world situations do face entrenched biases that may impact the effect of election aversion. For example, if our target context is a real-world competitive election, we must collect and adjust for treatment effect moderators that account for how the effect differs between the lab and real-world setting, such as baseline measures of expectations about gender bias.

G Relationship to Other Concepts

Here we clarify the relationship between our definition of the external validity and other concepts proposed in the literature.

Construct and ecological validity are important relevant concepts (Shadish, Cook and Campbell, 2002; Morton and Williams, 2010). Both help external validity, but they are not sufficient for external validity. *Construct validity* asks whether and how well experimental results speak to a theory of interest. Targets of the external validity analysis are often chosen based on a theory of interest, and thus, experiments with high construct validity are more likely to be externally valid. However, construct validity does not imply external validity. For example, as repeatedly found in the literature, small implementation differences in treatments, which are indistinguishable from a theoretical perspective, often induce a large variation in treatment effects. *Ecological validity*, also known as mundane experimental realism, asks "whether the methods, materials, and settings of the research are similar to a given target environment" (Morton and Williams, 2010). Again, experiments with high ecological validity are more likely to be externally valid because the targets of the external validity analysis are often chosen based on real-world settings. However, ecological validity might not be necessary or sufficient if, for example, the goal of the experiment is to test a formal model of strategic voting behavior.

Finally, we emphasize that concerns over external validity have a long history, and great scholars have introduced a variety of definitions for external validity. Thus, naturally, our definition of external validity cannot capture all conceptual and practical concerns raised in the literature. Notwithstanding the importance and utility of other definitions, we offer a definition of external validity based on the formal causal inference framework in Section 3.2, which admits coherent empirical approaches for external validity. The main goal of this paper is to develop this empirical approach for external validity.

References

Angrist, Joshua D. and William N. Evans. 1998. “Children and Their Parents’ Labor Supply: Evidence from Exogenous Variation in Family Size.” *The American Economic Review* 88(3):450–477.

Ansolabehere, Stephen and Brian F. Schaffner. 2017. “CCES Common Content, 2016.”. **URL:** *https://doi.org/10.7910/DVN/GDF6Z0*

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.

Bol, Damien. 2019. “Putting Politics in the Lab: A Review of Lab Experiments in Political Science.” *Government and Opposition* 54(1):167–190.

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.

Dickson, Eric S. 2011. Economics versus Psychology Experiments: Stylization, Incentives, and Deception. In *Cambridge Handbook of Experimental Political Science*, ed. James H. Kuklinski James N. Druckman, Donald P. Green and Arthur Lupia. New York, NY: Cambridge University Press chapter 5, pp. 58–72.

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.

Efron, Bradley and Robert J Tibshirani. 1994. *An Introduction To The Bootstrap.* CRC press.

Enamorado, Ted and Kosuke Imai. 2018. “CCES 2016 Voter Validation Supplemental Data.”. **URL:** *https://doi.org/10.7910/DVN/2NNA4L*

Gerber, Alan S and Donald P Green. 2012. *Field Experiments: Design, Analysis, and Interpretation*. WW Norton.

Hartman, Erin and F. Daniel Hidalgo. 2018. "An Equivalence Approach to Balance and Placebo Tests." *American Journal of Political Science* 62(4):1000–1013.
URL: *https://onlinelibrary.wiley.com/doi/abs/10.1111/ajps.12387*

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. "Bayesian Nonparametric Modeling for Causal Inference." *Journal of Computational and Graphical Statistics* 20(1):217–240.

Kanthak, Kristin and Jonathan Woon. 2015. "Women Don't Run? Election Aversion and Candidate Entry." *American Journal of Political Science* 59(3):595–612.

Morton, Rebecca B and Kenneth C Williams. 2010. *Experimental Political Science and the Study of Causality: From Nature to the Lab*. Cambridge University Press.

Rosenbaum, Paul R. 1984. "The Consequences of Adjustment For A Concomitant Variable That Has Been Affected by the Treatment." *Journal of the Royal Statistical Society: Series A (General)* 147(5):656–666.

Shadish, William R, Thomas D Cook and Donald T Campbell. 2002. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston: Houghton Mifflin.

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